



# STATISTICS IN TRANSITION

*new series*

---

An International Journal of the Polish Statistical Association and Statistics Poland

---

Okrasa W., Rozkrut D., Preface

## PART I

CELEBRATING 100<sup>TH</sup> ISSUE AND THE 30<sup>TH</sup> ANNIVERSARY

### Invited Paper

**Kalton G.**, Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day  
Comments and Discussions: **Pfeffermann D., Lehtonen R., Gershunskaya J., Lahiri P., Münnich R., C.,**  
**Wesołowski J.**, Rotation schemes and Chebyshev polynomials

## PART II

### Original Research Papers

**Ndlovu B. D., Melese S. F., Zewotir T.**, A nonparametric analysis of discrete time competing risks data:  
a comparison of the cause-specific-hazards approach and the vertical approach

**Vogt M., Lahiri P., Münnich R.**, Spatial Prediction in Small Area Estimation

**Tiwari K. K., Sharma V.**, Efficient estimation of population mean in the presence of non-response and  
measurement error

**Boumahdi M., Ouassou I., Rachdi M.**, Conditional density function for surrogate scalar response

**Öztaş Ayhan H.**, Models for survey nonresponse and bias adjustment techniques

**Białek J.**, Quality adjusted GEKS-type indices for price comparisons based on scanner data

**Olalude G. A., Yaya O. S., Olayinka H. A., Jimoh T. A., Adebisi A. A., Adesina O. A.**, Household expenditure  
in Africa: evidence of mean reversion

**Sharma A., Sharma V., Tokas S.**, Does economic freedom promote financial development? Evidence  
from EU countries

**Bhattacharjee A., Dey R.**, Bayesian modelling for semi-competing risks data in the presence of censoring

### Other articles

**Chugaievska S., Dehnel G., Targonskii A.**, Census administration in Ukraine: insight into the Polish experience  
in the context of international indicators analysis

**Grzenda W.**, Estimating the probability of leaving unemployment for older people in Poland using survival  
models with censored data

### Research Communicates and Letters

**Akhtar N., Khan S. A., Amin M., Khan A. A., Ali A., Manzoor S.**, Bayesian estimation of a geometric  
distribution using informative priors based on a Type-I censoring scheme

### Conference Announcement

**MET2023:** International Conference on Methodology of Statistical Research, 3–5 July, Warsaw, Poland

**30<sup>th</sup> Anniversary of the Statistics in Transition**

## EDITOR

Włodzimierz Okrasa *University of Cardinal Stefan Wyszyński, Warsaw and Statistics Poland, Warsaw, Poland*  
e-mail: w.okrasa@stat.gov.pl; phone number +48 22 – 608 30 66

## EDITORIAL BOARD

Dominik Rozkrut (Co-Chairman)	<i>Statistics Poland, Warsaw, Poland</i>
Waldemar Tarczyński (Co-Chairman)	<i>University of Szczecin, Szczecin, Poland</i>
Czesław Domański	<i>University of Łódź, Łódź, Poland</i>
Malay Ghosh	<i>University of Florida, Gainesville, USA</i>
Graham Kalton	<i>University of Maryland, College Park, USA</i>
Mirosław Krzysko	<i>Adam Mickiewicz University in Poznań, Poznań, Poland</i>
Partha Lahiri	<i>University of Maryland, College Park, USA</i>
Danny Pfeffermann	<i>Professor Emeritus, Hebrew University of Jerusalem, Jerusalem, Israel</i>
Carl-Erik Särndal	<i>Statistics Sweden, Stockholm, Sweden</i>
Jacek Wesolowski	<i>Statistics Poland, and Warsaw University of Technology, Warsaw, Poland</i>
Janusz L. Wywiłł	<i>University of Economics in Katowice, Katowice, Poland</i>

## ASSOCIATE EDITORS

Arup Banerji	<i>The World Bank, Washington, USA</i>	Andrzej Młodak	<i>Statistical Office Poznań, Poznań, Poland</i>
Misha V. Belkindas	<i>ODW Consulting, USA</i>	Colm A. O'Muirheartaigh	<i>University of Chicago, Chicago, USA</i>
Sanjay Chaudhuri	<i>National University of Singapore, Singapore</i>	Ralf Münnich	<i>University of Trier, Trier, Germany</i>
Eugeniusz Gatnar	<i>National Bank of Poland, Warsaw, Poland</i>	Oleksandr H. Osaulenko	<i>National Academy of Statistics, Accounting and Audit, Kiev, Ukraine</i>
Krzysztof Jajuga	<i>Wrocław University of Economics, Wrocław, Poland</i>	Viera Pacáková	<i>University of Pardubice, Pardubice, Czech Republic</i>
Alina Jędrzejczak	<i>University of Łódź, Poland</i>	Tomasz Panek	<i>Warsaw School of Economics, Warsaw, Poland</i>
Marianna Kotzeva	<i>EC, Eurostat, Luxembourg</i>	Mirosław Pawlak	<i>University of Manitoba, Winnipeg, Canada</i>
Marcin Kozak	<i>University of Information Technology and Management in Rzeszów, Rzeszów, Poland</i>	Mirosław Szreder	<i>University of Gdańsk, Gdańsk, Poland</i>
Danute Krapavickaite	<i>Institute of Mathematics and Informatics, Vilnius, Lithuania</i>	Imbi Traat	<i>University of Tartu, Tartu, Estonia</i>
Martins Liberts	<i>Bank of Latvia, Riga, Latvia</i>	Vijay Verma	<i>Siena University, Siena, Italy</i>
Risto Lehtonen	<i>University of Helsinki, Helsinki, Finland</i>	Gabriella Vukovich	<i>Hungarian Central Statistical Office, Budapest, Hungary</i>
Achille Lemmi	<i>Siena University, Siena, Italy</i>	Zhanjun Xing	<i>Shandong University, Shandong, China</i>

## EDITORIAL OFFICE

ISSN 1234-7655

Scientific Secretary

Marek Cierpiał-Wolan, *Statistical Office in Rzeszów, Rzeszów, Poland, e-mail: m.cierpial-wolan@stat.gov.pl*

Managing Editor

Adriana Nowakowska, *Statistics Poland, Warsaw, e-mail: a.nowakowska3@stat.gov.pl*

Secretary

Patryk Barszcz, *Statistics Poland, Warsaw, Poland, e-mail: p.barszcz@stat.gov.pl, phone number +48 22 – 608 33 66*

Technical Assistant

Rajmund Litkowiec, *Statistical Office in Rzeszów, Rzeszów, Poland, e-mail: r.litkowiec@stat.gov.pl*

© Copyright by Polish Statistical Association, Statistics Poland and the authors, some rights reserved. CC BY-SA 4.0 licence



## Address for correspondence

*Statistics Poland, al. Niepodległości 208, 00-925 Warsaw, Poland, tel./fax: +48 22 – 825 03 95*

## Preface

This issue is the hundredth in 30 years of publishing *Statistics in Transition*. The first issue appeared in July 1993, and for the next fifteen years it was a semi-annual publication. In 2007 the title of the journal was slightly changed to *Statistics in Transition new series* and it became a quarterly publication. To celebrate the historical significance of these milestones, we dedicate the first part of this issue to them, opening it with a specially prepared Invitation Paper, along with four discussion pieces of the issues raised in that paper.

With a sense of deep gratitude and the highest appreciation we would like to thank, both personally and on behalf of all the editorial bodies, Professor Graham Kalton for preparing his Invited Paper entitled ***Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day***. Dr. Kalton is a long-time friend of our journal, and he serves as a member of our Editorial Board. The issues discussed in Dr. Kalton's paper are particularly appropriate at this time as major changes are taking place in survey research methods and in sources of official statistics. The paper and the discussion pieces should therefore be of interest to members of the international statistician community and to members of national statistical offices.

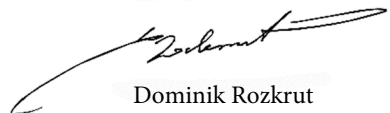
Despite the relatively short time for reactions, we are grateful to five eminent experts, four of whom are associated with *SiTns*, for preparing four discussion pieces related to the paper. The authors of the four discussions are Professor Danny Pfeffermann, Dr. Julie Gershunskaya and Professor Partha Lahiri, Professor Risto Lehtonen, and Professor Ralf Münnich. Each of the discussions provides insightful observations supplementing some of the issues picked out from those discussed by Graham Kalton. They share concerns about the current challenges to probability sampling and design-based inference primarily caused by the serious declines in response rates, especially in high-income countries. They point to the possibilities of using alternative modalities (administrative data, big data, internet data, scientific data, etc.) for data collection that can supplement or replace probability samples. They describe the considerable body of research that is in progress to enable these alternative data sources to produce valid population estimates from the nonprobability samples associated with the modalities, and to the data integration methods that are being developed to combine the data obtained from different sources.

An *addendum* to this section contains a paper by Professor Jacek Wesolowski entitled *Rotation schemes and Chebyshev polynomials*, as being inspired in a way by the Invited Paper, and as an indication of other types of effects that it may have as well.

It is noteworthy that as our journal celebrates its 30<sup>th</sup> anniversary, the journal's name *Statistics in Transition* well reflects the radical changes in the methodology of survey statistics and official statistics that are currently underway, as indicated in the Invited Paper and the discussions in this section.



Włodzimierz Okrasa  
Editor, *Statistics in Transition new series*



Dominik Rozkrut  
President, Statistics Poland



## CONTENTS

<b>Okrasa W., Rozkrut D.</b> , Preface .....	I
From the Editor .....	IX

## PART I

CELEBRATING 100<sup>TH</sup> ISSUE AND THE 30<sup>TH</sup> ANNIVERSARY

## Invited Paper

<b>Kalton G.</b> , Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day.....	1
<b>Pfeffermann D.</b> , Comments .....	23
<b>Lehtonen R.</b> , Comments .....	27
<b>Gershunskaya J., Lahiri P.</b> , Discussion .....	31
<b>Münnich R., C.</b> , Discussion .....	39
<b>Kalton G.</b> , Rejoinder .....	43
<b>Wesołowski J.</b> , Rotation schemes and Chebyshev polynomials .....	47

## PART II

## Original Research Papers

<b>Ndlovu B. D., Melese S. F., Zewotir T.</b> , A nonparametric analysis of discrete time competing risks data: a comparison of the cause-specific-hazards approach and the vertical approach .....	61
<b>Vogt M., Lahiri P., Münnich R.</b> , Spatial Prediction in Small Area Estimation .....	77
<b>Tiwari K. K., Sharma V.</b> , Efficient estimation of population mean in the presence of non-response and measurement error .....	95
<b>Boumahdi M., Ouassou I., Rachdi M.</b> , Conditional density function for surrogate scalar response .....	117
<b>Öztaş Ayhan H.</b> , Models for survey nonresponse and bias adjustment techniques .....	139
<b>Białek J.</b> , Quality adjusted GEKS-type indices for price comparisons based on scanner data ....	151
<b>Olalude G. A., Yaya O. S., Olayinka H. A., Jimoh T. A., Adebisi A. A., Adesina O. A.</b> , Household expenditure in Africa: evidence of mean reversion .....	171
<b>Sharma A., Sharma V., Tokas S.</b> , Does economic freedom promote financial development? Evidence from EU countries .....	187
<b>Bhattacharjee A., Dey R.</b> , Bayesian modelling for semi-competing risks data in the presence of censoring .....	201

## Other articles

*XL Multivariate Statistical Analysis (MAS 2022 Conference), Lodz, Poland.*

<b>Chugaievska S., Dehnel G., Targonskii A.</b> , Census administration in Ukraine: insight into the Polish experience in the context of international indicators analysis .....	213
--	-----

*XLI Scientific Conference of the Classification and Data Analysis Section (SKAD 2022)*

<b>Grzenda W.</b> , Estimating the probability of leaving unemployment for older people in Poland using survival models with censored data .....	241
--	-----

## Research Communicates and Letters

<b>Akhtar N., Khan S. A., Amin M., Khan A. A., Ali A., Manzoor S.</b> , Bayesian estimation of a geometric distribution using informative priors based on a Type-I censoring scheme .....	257
---	-----

## Conference Announcement

<b>MET2023:</b> International Conference on Methodology of Statistical Research, will be held on 3–5 July in Warsaw, Poland.....	265
About the Authors .....	267



## Submission information for Authors

*Statistics in Transition new series (SiTns)* is an international journal published jointly by the Polish Statistical Association (PTS) and Statistics Poland, on a quarterly basis (during 1993–2006 it was issued twice and since 2006 three times a year). Also, it has extended its scope of interest beyond its originally primary focus on statistical issues pertinent to transition from centrally planned to a market-oriented economy through embracing questions related to systemic transformations of and within the national statistical systems, world-wide.

The *SiTns* seeks contributors that address the full range of problems involved in data production, data dissemination and utilization, providing international community of statisticians and users – including researchers, teachers, policy makers and the general public – with a platform for exchange of ideas and for sharing best practices in all areas of the development of statistics.

Accordingly, articles dealing with any topics of statistics and its advancement – as either a scientific domain (new research and data analysis methods) or as a domain of informational infrastructure of the economy, society and the state – are appropriate for *Statistics in Transition new series*.

Demonstration of the role played by statistical research and data in economic growth and social progress (both locally and globally), including better-informed decisions and greater participation of citizens, are of particular interest.

Each paper submitted by prospective authors are peer reviewed by internationally recognized experts, who are guided in their decisions about the publication by criteria of originality and overall quality, including its content and form, and of potential interest to readers (esp. professionals).

Manuscript should be submitted electronically to the Editor:

sit@stat.gov.pl,

GUS/Statistics Poland,

Al. Niepodległości 208, R. 296, 00-925 Warsaw, Poland

It is assumed, that the submitted manuscript has not been published previously and that it is not under review elsewhere. It should include an abstract (of not more than 1600 characters, including spaces). Inquiries concerning the submitted manuscript, its current status etc., should be directed to the Editor by email, address above, or w.okrasa@stat.gov.pl.

For other aspects of editorial policies and procedures see the *SiT* Guidelines on its Web site: <https://sit.stat.gov.pl/ForAuthors>.





*STATISTICS IN TRANSITION* new series, June 2023

Vol. 24, No. 3, pp. VII–VIII

## **Policy Statement**

The broad objective of *Statistics in Transition new series* is to advance the statistical and associated methods used primarily by statistical agencies and other research institutions. To meet that objective, the journal encompasses a wide range of topics in statistical design and analysis, including survey methodology and survey sampling, census methodology, statistical uses of administrative data sources, estimation methods, economic and demographic studies, and novel methods of analysis of socio-economic and population data. With its focus on innovative methods that address practical problems, the journal favours papers that report new methods accompanied by real-life applications. Authoritative review papers on important problems faced by statisticians in agencies and academia also fall within the journal's scope.

\*\*\*

## Abstracting and Indexing Databases

*Statistics in Transition new series* is currently covered in:

BASE – Bielefeld Academic Search Engine	JournalGuide
CEEOL – Central and Eastern European Online Library	JournalTOCs
CEJSH (The Central European Journal of Social Sciences and Humanities)	Keepers Registry
CNKI Scholar (China National Knowledge Infrastructure)	MIAR
CNPIEC – cnpLINKer	Microsoft Academic
CORE	OpenAIRE
Current Index to Statistics	ProQuest – Summon
Dimensions	Publons
DOAJ (Directory of Open Access Journals)	QOAM (Quality Open Access Market)
EconPapers	ReadCube
EconStore	RePec
Electronic Journals Library	SCImago Journal & Country Rank
Elsevier – Scopus	TDNet
ERIH PLUS (European Reference Index for the Humanities and Social Sciences)	Technische Informationsbibliothek (TIB) – German National Library of Science and Technology
Genamics JournalSeek	Ulrichsweb & Ulrich’s Periodicals Directory
Google Scholar	WanFang Data
Index Copernicus	WorldCat (OCLC)
J-Gate	Zenodo

## From the Editor

### Part II

This part contains a set of twelve articles written by thirty-four authors, indicating a dominant collaborative form of authorship (almost three authors per article on average) and showing a trend towards that observed in the natural and technical sciences. In this regard, another positive feature appears, that is the growing scope of internationalization of our journal. This is also important for the wide variety of its input and for the appropriate reach on the readership side, which we care so much about. In fact, the authors of the texts in this part come from thirteen countries: South Africa, Germany, USA, India, Morocco, France, Turkey, Poland, Nigeria, United Kingdom, Ukraine, Pakistan, and Sweden. Also, the spectrum of topics covered in these articles is suitably diverse, from theoretical interests to applications, but with concern shared about the usefulness of even abstract approaches to solving real-world problems in an innovative way.

### Research articles

In the first paper, prepared by **Bonginkosi Duncan Ndlovu**, **Sileshi Fanta Melesse**, and **Temesgen Zewotir** titled *A nonparametric analysis of discrete time competing risks data: a comparison of the cause-specific-hazards approach and the vertical approach* the vertical model as a nonparametric model for analysis of discrete time competing risks data was presented. The secondary objective of this article is to compare the proposed model to this model. The authors pay particular attention to the estimates for the cause-specific-hazards and the cumulative incidence functions as well as their respective standard errors. It was shown that the standard errors for the estimates of these quantities were identical under both models. It is a roundabout way of estimating the cause-specific-hazards, however, there are cases in practice where these quantities cannot be estimated directly from the data such as when some of the subjects have failed with unknown failure causes. Furthermore, the cause-specific-hazards are not appropriate for application in the presence of a sizable proportion of cured subjects. The cause-specific-hazards model cannot handle these data complications. The proposed model, therefore, offers a possibility that the proposed model can also be upscaled to handle these challenges in discrete time.

**Martin Vogt, Partha Lahiri, and Ralf Münnich** in their article *Spatial Prediction in Small Area Estimation* have developed a hierarchical Bayes methodology for an extension of the well-celebrated Fay-Herriot model that incorporates spatial correlation using an intrinsic CAR model, and proved the propriety of the posterior distribution for our proposed model. The authors have tested the effect of covariates on the estimation results. An application to SAIPE data revealed that modeling spatial correlation can considerably improve on the associated hierarchical Bayes methodology if the area-specific auxiliary data are either weak. Small area estimation methods have become a widely used tool to provide accurate estimates for regional indicators such as poverty measures. Recent research has provided evidence that spatial modeling still can improve the precision of regional and local estimates.

In the next manuscript *Efficient estimation of population mean in the presence of non-response and measurement error* **Kuldeep Kumar Tiwari and Vishwantra Sharma** have considered ten estimators of population mean and studied them in the context of non-response and measurement error. In real-world surveys, non-response and measurement errors are common, therefore studying them together seems rational. Some population mean estimators are modified and studied in the presence of non-response and measurement errors. Bias and mean squared error expressions are derived under different cases. For all estimators, a theoretical comparison is made with the sample mean per unit estimator. The Monte-Carlo simulation is used to present a detailed picture of all estimators' performance. The expressions for bias and MSE for all the estimators in various cases were obtained.

**Mounir Boumahdi's, Idir Ouassou's, and Mustapha Rachdi's** paper *Conditional density function for surrogate scalar response* presents the estimator of the conditional density function of surrogated scalar response variable given a functional random one. A conditional density function by using the available (true) response data and the surrogate data was constructed, and some asymptotic properties of the constructed estimator in terms of the almost complete convergences were built. As a result, the authors have compared the estimator with the classical estimator through the Relative Mean Square Errors (RMSE). Finally, this analysis by displaying the superiority of the authors' estimator in terms of prediction when one is lacking complete data was completed. In this paper the almost complete convergence of conditional density function for surrogated scalar response variable given a functional random by using validation sample set was presented, and the performance of the estimator  $\hat{f}_{xR}(y)$  than  $\hat{f}_{xV}(y)$  to reduce RMSE by using the simulated data was shown.

The article by **H. Öztaş Ayhan** entitled *Models for survey nonresponse and bias adjustment techniques* discusses the aspects and the complex nature of the

nonresponse in sample surveys. An overview of the components of the bias due to nonresponse was performed. The survey unit nonresponse bias has been examined alternatively by taking response amounts which are fixed initially and also by taking the response amounts as random variables. Nonresponse bias components were illustrated for each alternative approach and the amount of bias was computed for each case. The evaluation of the nonresponse bias as nonresponse error or nonresponse rate was misleading. The nonresponse bias may seem to be related to the response rates for a given study. Increasing response rate may not always correspond to decreasing nonresponse bias for a given study. This paper has shown alternative approaches to nonresponse bias. In addition to this, the causes of the nonresponse bias can also be obtained from empirical studies of components and models relating to the covariates of survey participation and non-participation. The current research examined the response amounts as fixed initially. The proposed methodology has shown the effect of bias of nonresponse, which is based as the product of “amount of nonresponse rate” and the “difference between the response and nonresponse strata means”.

**Jacek Białek's** paper *Quality adjusted GEKS-type indices for price comparisons scanner data* deals with the two new multilateral indices, the idea of which resembles the GEKS method, but which perform additional quality adjustment and deviate from the classical approach in which the base formula of the GEKS index is a superlative index. The empirical analysis has confirmed that the two proposed indices (GEKS-AQU and GEKS-AQI) satisfy most of commonly accepted tests for multilateral indices including the identity test. The study has shown that differences between the proposed indices and other considered multilateral indices appear only with large variability of quantity in homogeneous groups of products. It should be noted that quite surprisingly, the price volatility did not play a significant role in the empirical study as determinants of differences between multilateral indices. The same study has also shown that the computation time needed in the case of the GEKS-AQU and GEKS-AQI indices is average compared to most other multilateral indices. The previously known multilateral indices (Geary-Khamis, GEKS, TPD, CCDI, and SPQ) as well as the new indices proposed and discussed in the paper (GEK-AQU, GEKS-AQI, and their weighted versions: WGEKS-AQU and WGEKS-AQI) are implemented in the PriceIndices R package, and thus the reader can verify their usefulness on their own data sets.

**Gbenga A. Olalude, OlaOluwa S. Yaya, Hamed A. Olayinka, Toheeb A. Jimoh, Aliu A. Adebisi, and Oluwaseun A. Adesina** in their paper *Household expenditure in Africa: evidence of mean reversion* investigate the mean reversion in household consumption expenditure in 38 African countries. The expenditure series used were

the percentage of nominal Gross Domestic Product (GDP), each spanning 1990 to 2018. Due to a small sample size of time series of household expenditure, with possible structural breaks, the authors used the Fourier unit root test approach, which enabled modeling both smooth and instantaneous breaks in the expenditure series. The results showed non-mean reversion in the consumption expenditure pattern of Egypt, Madagascar and Tunisia, while mean reversion was detected in the remaining 35 countries. Thus, the majority of African countries are on the verge of recession once shocks that affect the growth of GDP are triggered. Findings in this paper are of relevance to policymakers on poverty alleviation programmes in those selected countries.

In the paper *Does economic freedom promote financial development? Evidence from EU countries* Anand Sharma, Vipin Sharma, and Shekhar Tokas empirically explore the relationship between economic freedom and financial development in EU countries. Using panel data covering the years 2000–2017 and employing fixed effects, random effects, and the generalised method of moments (GMM), the paper examines the effect of economic freedom on financial development. The research results demonstrate that greater economic freedom is conducive to financial development in the EU. These findings remain robust to the use of an alternative index of economic freedom. The results imply that policies which promote economic freedom are likely to raise the level of a country's financial development. The article uses an index of overall financial development as a dependent variable and does not focus on the financial markets and financial institutions sub-indices. Future research may attempt to consider the effect of economic freedom on the development of financial markets and financial institutions.

*Bayesian modelling for semi-competing risks data in the presence of censoring* prepared by Atanu Bhattacharjee and Rajashree Dey presents the semi-competing risks framework as a way of investigating variation in risk for a non-terminal event where the occurrence of the event is subject to a terminal event. In this context, the authors have analyzed the semi-competing risk data using the proposed AFT illness death model, which serves as a helpful complement of the traditional hazard-based model of say. The work is dedicated to overcoming the existing challenges by the applications of R programming and data illustration. The authors arrived at a conclusion that the developed methods are suitable to run and easy to implement in R software. The selection of covariates in the AFT model can be evaluated using model selection criteria such as the Deviance Information Criteria (DIC) and Log-pseudo marginal likelihood (LPML). Various extensions of the AFT model, such as AFT-DPM and AFT-LN, have been demonstrated. The final model was selected based on minimum DIC values and larger LPML values.

### **Other articles**

**Svitlana Chugaievska's, Grażyna Dehnel's, and Andrey Targonskii's** paper *Census administration in Ukraine: insight into the Polish experience in the context of international indicators analysis* deals with the analysis of a number of international indices that are relevant for respondent participation in statistical surveys, and particularly in the context of the next population census. Three groups of indices were identified: indicators of electronic document circulation, indicators of sustainable economic development, and social indicators. Considering each of these indices, the situation of Poland is significantly better compared to that of Ukraine, where the last national census was conducted only once in 2001. A comparative analysis of census questionnaires used in Poland in 2021 and in Ukraine in 2019 revealed that the Polish census form was not only longer (73 vs. 50 questions), but also included some aspects that were absent from the Ukrainian questionnaire, e.g. a section about family ties in the household. As regards respondent participation, a very low percentage of young respondents self-enumerated online, probably because of insufficient information about how to use the web application.

The article by **Wioletta Grzenda** entitled *Estimating the probability of leaving unemployment for older people in Poland using survival models with censored data* assesses the probability of leaving unemployment for people aged 50–71 based on their characteristics and the length of the unemployment period. The data from the Labour Force Survey for 2019–2020 were used. The key factors determining employment status are identified using the proportional hazard model. The author takes these factors into account and uses the direct adjusted survival curve to show how the probability of returning to work in Poland changes as people age. Due to the fact that not many people take up employment around their retirement age, an in-depth evaluation of the accuracy of predictions obtained via the models is crucial to assess the results. Hence, in this paper, a time-dependent ROC curve is used. The results indicate that the key factor that influences the return to work after an unemployment period in the case of older people in Poland is whether they reached the age of 60. Other factors that proved important in this context are the sex and the education level of older people.

### **Research Communicates and Letters**

The *research communicates and letters* includes the paper by **Nadeem Akhtar, Sajjad Ahamad Khan, Muhammad Amin, Akbar Ali Khan, Amjad Ali, and Sadaf Manzoor**, entitled *Bayesian estimation of a geometric distribution using informative priors based on a Type-I censoring scheme*. The authors discuss the geometric distribution parameter that is estimated under a type-I censoring scheme by means of

the Bayesian estimation approach. The Beta and Kumaraswamy informative priors, as well as five loss functions are used for this purpose. Expressions of Bayes estimators and Bayes risks are derived under the Squared Error Loss Function (SELF), the Quadratic Loss Function (QLF), the Precautionary Loss Function (PLF), the Simple Asymmetric Precautionary Loss Function (SAPLF), and the DeGroot Loss Function (DLF) using the two aforementioned priors. The prior densities are obtained through prior predictive distributions. Simulation studies are carried out to make comparisons using Bayes risks. Finally, a real-life data example is used to verify the model's efficiency. An extensive simulation study and a real-life data analysis is employed to validate the importance of the proposed strategy. The numerical results reveal that Beta is an appropriate prior and SELF is a better loss function while analysing discrete geometric life testing model under type-I censoring scheme. The real-life data analysis cements these findings.

**Włodzimierz Okrasa**

Editor

© Włodzimierz Okrasa. Article available under the CC BY-SA 4.0 licence





# Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day

Graham Kalton<sup>1</sup>

## Abstract

At the beginning of the 20th century, there was an active debate about random selection of units versus purposive selection of groups of units for survey samples. Neyman's (1934) paper tilted the balance strongly towards varieties of probability sampling combined with design-based inference, and most national statistical offices have adopted this method for their major surveys. However, nonprobability sampling has remained in widespread use in many areas of application, and over time there have been challenges to the Neyman paradigm. In recent years, the balance has tilted towards greater use of nonprobability sampling for several reasons, including: the growing imperfections and costs in applying probability sample designs; the emergence of the internet and other sources for obtaining survey data from very large samples at low cost and at high speed; and the current ability to apply advanced methods for calibrating nonprobability samples to conform to external population controls. This paper presents an overview of the history of the use of probability and nonprobability sampling from the birth of survey sampling at the time of A. N. Kiær (1895) to the present day.

**Key words:** Anders Kiær, Jerzy Neyman, representative sampling, quota sampling, hard-to-survey populations, model-dependent inference, internet surveys, big data, administrative records.

## 1. Introduction

This paper presents a selection of the major developments that have taken place over the years since social surveys were first introduced in the late 19th century. I restrict my coverage to surveys of households and persons and my focus is on the sampling methods used to conduct such surveys. Major changes have also taken place in modes of data collection, in questionnaire design, and in other aspects of survey research over the years, but these topics are outside the scope of this paper. My paper on the more general theme of survey research over the past 60 years overlaps with this paper and gives greater coverage on some topics (Kalton, 2019).

---

<sup>1</sup> Joint Program in Survey Methodology, University of Maryland, College Park, MD, USA.  
E-mail: [gkalton@gmail.com](mailto:gkalton@gmail.com). ORCID: <https://orcid.org/0000-0002-9685-2616>.

The changes that have occurred in methods of survey sampling have arisen for many reasons, including developments in sampling theory, the continuing growth in computer power (that was non-existent for the first fifty years of survey research), new sampling frames, and the problems created by a broader and more challenging range of applications of social surveys that has occurred as the potential for survey research has been more fully recognized. While acknowledging these changes, it is noteworthy that many aspects of the sampling methods that have been superseded over time have remained relevant. Indeed, much of the current discussion of the use of nonprobability sampling and big data sources has roots in the early days of survey research.

Without attempting to date the origins of survey research, early applications of survey research for studying the social conditions of populations took off in the late 1800's. English examples include Charles Booth's large-scale survey of the social conditions of the population of London that was started in 1886, Seebohm Rowntree's survey of working-class poverty in York that was conducted a decade later, and Bowley's survey of working-class conditions in Reading in 1912, which he followed up with surveys in four other English towns (three of which were conducted by Burnett-Hurst under Bowley's direction). See Caradog Jones (1949) for the early surveys in England, Converse (2017) for an account of the history of survey research in the United States from its beginnings at the turn of the century through until 1960, and Stephan (1948) for a history of the use of sampling procedures dating back from earlier times through until the 1940's, primarily in the United States.

The London and York surveys were complete censuses of the surveys' target populations. As complete censuses, they were deemed statistically acceptable at the time; they were known as 'monographs' of their local communities. For the London survey, the target population was households with school-aged children, while for the York survey it was households that did not have servants (conducted only in streets that were likely to contain households without servants). Bowley had long argued for the use of sampling for such surveys, and he played a major role in its adoption (Aldrich, 2008). He used sampling for the first time in the five towns surveys, where systematic sampling was employed (Bowley, 1913), and he introduced the idea of measuring sampling errors for survey estimates.

As Kish (1995) notes, the emergence of the field of survey sampling can be dated from work led by the Norwegian statistician Anders Kiær, the first Director of Statistics Norway. Kiær developed a sampling method that he termed "representative sampling". Kiær's method of purposive sampling is worth reviewing both for the procedures he devised to make a sample nationally 'representative' and for the reactions to the method from statisticians attending meetings of the International Statistical Institute (ISI) at the time. The next section provides a brief overview of these issues.

## 2. Kiær's Representative Method of Statistical Surveys

Kiær's sampling methodology is described in detail in his monograph *The Representative Method of Statistical Surveys*, first published in Norwegian in 1897 and republished in 1976 with an English translation (Kiær, 1976). The monograph provides a good deal of detail on the sample designs Kiær developed for two large-scale surveys—one on personal income and property (PIP) and the other on living conditions (LC)—as well as reporting the objections to his methods that he received when he presented them at ISI meetings. As distinct from the surveys of English towns cited above, Kiær aimed to produce survey estimates for the whole of Norway. For this purpose, he developed two-stage area sample designs for his surveys: at the first stage, he selected a “representative” sample of administrative districts (rural districts or counties, towns, and cities); at the second stage, he drew samples of people for each survey. The choice of the sampled first-stage units was carefully fashioned to give geographical spread and to achieve a good representation of the Norwegian population in terms of characteristics collected in the 1891 Population Census (e.g., age, marital status, occupation, urbanicity).

The sample for the PIP survey was defined as men aged 17, 22, 27, etc. who had names starting with certain letters, selected from 1891 census records that were being processed at the time, with a total sample size of around 11,400 men. The sample size for the LC survey was around 80,000 adults. The sample size to be obtained in each selected rural county was specified based on calculations from census data; within selected counties, the enumerators were instructed to follow certain routes and to select different types of houses, but otherwise they were left to make the selections. In the smaller towns, every 9<sup>th</sup>, 5<sup>th</sup>, or 3<sup>rd</sup> house was selected. An extra sampling stage was introduced in the largest towns. For example, the sample of houses in Oslo was selected within a sample of streets. Moreover, a higher proportion of the streets with larger populations was included in the sample, but this feature was counterbalanced by the selection of houses at a lower rate in the large streets.

The driving objective with Kiær's approach was to produce a representative sample that would constitute a microcosm of the Norwegian population. He invented some intricate methods to attempt to achieve this objective. His purposive selection of first stage administrative units sometimes incorporated ideas of probability proportional to size sampling and subsampling at different rates in compensation, thereby avoiding an excessive sample concentration in a few large districts. Similarly, his street sample in Oslo has the same feature. He also employed a simple 2:1 weighting adjustment to compensate for the smaller proportion of members of the rural population in the PIP survey. (Before the advent of computers, anything other than simple integer weighting adjustments would have been extremely difficult to routinely apply.)

Despite his thoughtful approach, Kiær encountered a great deal of criticism of his methods when he presented them to the ISI in 1895. The dominant criticism, however, was not of the representative method, *per se*, but rather of a sample-based enquiry rather than a complete enumeration. In the words of one strong critic, von Mayr: “We remain firm and say: no calculations when observations can be made”. Kiær also made presentations on the representative method at the 1897, 1901, and 1903 ISI sessions, at which they were subjected to similar criticisms, together with another one. At the 1903 session, von Bortkiewicz reported the results of a significance test he had conducted that found that Kiær’s representative samples were not truly representative. See Kruskal and Mosteller (1980) for a detailed account of the ISI sessions.

At the same time, Kiær expertise was under attack at home for the LC survey, which was conducted on behalf of a parliamentary labor commission to inform a very contentious social security act that would provide highly expensive disability insurance. A three-person “critique committee” was established to review the commission’s major recommendation and its statistical basis. One committee member, the actuary Jens Hjorth, was extremely critical of Kiær’s statistics, including the survey design, the representative sample design, and the analysis. The attacks on the statistics that Kiær’s produced for the commission were forceful, extensive, and widely debated. In the end, based on the results of some new surveys, Kiær admitted that he had initially seriously underestimated the extent of disability. After that time, representative sampling for large-scale surveys disappeared in Norway. Lie (2002) provides an informative account of the rise and fall of Kiær’s representative sampling method.

The ISI discussion of survey sampling fell into abeyance until 1924 when the ISI appointed a commission for studying the application of the representative method in statistics. By that time, the idea of a “partial investigation” was widely accepted. In its 1926 report (Jensen, 1926), the Commission concluded that a sample was acceptable if it was sufficiently representative of the whole. To satisfy this condition the sample could be produced either by random selection with equal probability or by purposive selection of groups with a representative overall sample. The report also recommended that the survey results should, wherever possible, be accompanied by an indication of the errors to which they are liable.

### 3. Neyman’s Seminal Paper

In 1934, Neyman presented his classic paper comparing the methods of random and purposive selection to the Royal Statistical Society (Neyman, 1934). Covering more than the comparison, the paper contained a detailed discussion of a methodology for making inferences from random—or, more generally, probability—samples of finite populations, including providing a definition of a confidence interval in this context.

He also critically examined the assumptions made when using data from a purposive sample to produce an accurate estimate of a population parameter.

He discussed the sample design of purposive selection of groups used by Gini and Galvani in selecting a sample of records from the already-processed Italian General Census of 1921 that was to be used as the basis for later analysis. For their sample, Gini and Galvani (1929) selected a sample of twenty-nine of the 214 districts in Italy, balanced on seven covariables (note that departs from Kiær's stipulation that a large wide-spread sample of areas is needed). While the sample worked well for the averages of the control variables, it often failed to adequately represent the national population for other characteristics, and for the distributions of the control variables. These findings led them to raise questions about representative sampling.

Neyman's paper was a watershed for survey sampling, leading to widespread adoption of probability sampling, particularly by national statistical offices. It also led to the development of an extensive range of sampling methods and the associated theory applicable to a variety of practical survey problems, as described in the several texts on survey sampling that appeared in the 1950's. The many contributions of statisticians at the U.S. Census Bureau led by Morris Hansen are particularly noteworthy; see, for example, the two-volume text by Hansen, Hurwitz, and Madow (1953). Statisticians active in research on sample designs for agricultural surveys, such as Yates in England and Mahalanobis in India, also made important contributions to the advancement of the subject. The sampling text by Yates (1949) was among the first books on survey sampling methods. In 1950, Mahalanobis went on to establish and lead the famous socio-economic National Sample Survey (NSS) of India. An interesting feature of the NSS sample design was that the sample was composed of four replicate samples. The survey results were presented for each replicate separately as well as for the full sample, with the aim of communicating to readers an indication of the amount of sampling error in the survey estimates (see, for example, Mahalanobis, 1946). This was thus a forerunner of variance estimation using replication methods.

Note that perfect application of Neyman's design-based inference for probability sampling depends on:

- The availability of a sampling frame that provides complete coverage of the finite target population;
- A sample design that assigns known and non-zero selection probabilities to every element in the target population;
- Survey responses from every sampled unit; and
- The use of survey weights in the analysis to compensate for unequal selection probabilities.

Under these conditions (and assuming no response errors), survey estimates can be computed that are design-consistent estimates of the population parameters without the need to make any assumptions about the characteristics of the survey population. Model assumptions made about the population structure may be used to make the sample design more efficient or in the computation of the survey estimates, but the consistency of the survey estimates remains irrespective of the validity of the model. What the model assumptions do affect is the precision of the survey estimates. For example, in a stratified sample, if the sampling fraction in a stratum is set at a higher rate because the elements in a stratum are incorrectly modeled to be more variable, the (weighted) sample mean will still be unbiased, but it will be less precise than if the stratum element variance has been correctly modeled. Similarly, if a set of auxiliary variables  $\mathbf{X}$  is available for all population elements, and a function of the  $\mathbf{x}$ 's,  $f(\mathbf{X})$ , is used as a working model to predict the survey variable  $y$ , then the finite population total may be estimated by

$$\hat{Y}_d = \sum_U \hat{f}(\mathbf{X}_i) + \sum_S w_i e_i, \quad (1)$$

where  $\sum_U$  and  $\sum_S$  denote summations over the population and sample respectively,  $\hat{f}(\mathbf{X}_i)$  denotes the model estimate of  $y_i$  using the sample estimates of the unknown model parameters,  $e_i = y_i - \hat{f}(\mathbf{X}_i)$ , and the weight  $w_i$  is the inverse of element  $i$ 's selection probability. By including the weighted estimate of the population total of the  $e_i$ 's in this estimate,  $\hat{Y}_d$  is a consistent estimator of the population total  $Y$  irrespective of the suitability of the working model; the choice of working model affects only the precision of the estimate  $\hat{Y}_d$ . This estimator is model-assisted, using the terminology coined by Särndal, Swensson, and Wretman (1992), but it is not model-dependent. For simple random sampling, Cochran (1953) gave an early example of a model-assisted estimator with the ratio estimator  $\hat{Y} = (\bar{y}/\bar{x})X$ , where  $X$  denotes the population total for the auxiliary variable  $x$ . An additional, important, feature of design-based inference is that estimates of the variances of sample estimates can be computed from the sample itself.

While the lack of dependence of design-based inference on model assumptions is the major attraction of probability sampling, it needs to be acknowledged that probability sampling is rarely perfectly executed in practice. There are two main sources of imperfection: noncoverage and nonresponse. Noncoverage, which arises because the sampling frame fails to include some elements of the target population, is widespread and its magnitude is often underrated. Area sampling is widely used in social surveys, selecting a probability sample of geographical areas, listing the households or dwelling units in the sampled areas, selecting a probability sample of households, and selecting either all or a probability sample of persons in those households. Even when the sample

of areas provides complete geographical coverage, noncoverage arises often from incomplete listing of households or dwelling units within sampled areas, and from incomplete listing of persons within sampled households. Nonresponse occurs when a sampled element fails to provide acceptable responses to some or all the survey questions. In the early years of probability sampling, response rates were high, and these two sources of imperfection were treated as minor blemishes that received little attention. They were either ignored or treated by simple weighting adjustments (simple, in part because more complex adjustments were computationally infeasible at the time).

Probability sampling has two main drawbacks to be balanced against the theoretical attractions of design-based inference: cost and timeliness. The extra costs of probability sampling include the costs of tracking down sampled individuals, including repeat calls when the individual is not initially available. When area sampling is used, the sampling costs also include the costs of listing units within sampled areas. For similar reasons, collecting survey data from a probability sample takes longer, making the production of the survey estimates less timely. Timeliness is important for all surveys, but particularly for surveys where the results are highly time-dependent, such as political polls, surveys of outbreaks of certain infections, and surveys of areas that have experienced a recent disaster.

A variety of less rigorous sampling methods are used in an attempt to apply a probability sampling approach to address these drawbacks. However, since all these methods require modeling assumptions, none of them can be classified as probability sampling. For convenience, they are called ‘pseudo-probability’ methods in what follows. In the early days of design-based inference, the quasi-probability sampling method known as quota sampling was widely used in market research and in other applications. That method is described in Section 4. Three other quasi-probability sampling methods are described briefly in Section 5.

#### **4. Quota Sampling**

To set the scene for the need for imposing quota controls on a sample of the general population, consider the infamous Literary Digest Poll of 1936. To forecast the outcome of the 1936 U.S. Presidential Election, the Literary Digest mailed a questionnaire to a sample of ten million individuals selected from telephone directories, lists of automobile owners, and registered voters. The results obtained from the two million respondents indicated a clear-cut victory for Alf Landon with 57 percent of the vote, whereas in fact Franklin Roosevelt won with 61 percent of the vote. The upper-class bias of the sample, and of the respondents within the sample, is a major part of the explanation of the discrepancy between these percentages. No weighting adjustments

were employed to attempt to address the bias at the time. (Lohr and Brick, 2017, reweighted the sample using respondents' reports of their voting in the 1932 election, and these adjustments led to a correct prediction of the outcome, but the estimate of the vote for Roosevelt still fell far short of the actual vote.) This study serves to demonstrate that a large sample size does not necessarily yield good estimates. See Converse (2017) for more details.

Market researchers and pollsters developed the methods of quota sampling separately from the developments in probability sampling, with the aim of addressing the biases from uncontrolled sampling. There are various forms of quota sampling, with the essence of all of them being to control the types of persons to be interviewed. Interviewers are instructed to make their samples of respondents conform to specified quota controls by such characteristics as sex, age group, and employment status. The controls could be independent (e.g., so many men and so many women, so many persons over 35 and so many persons 35 years of age or less) or the numbers to be interviewed could be interrelated (e.g., so many men over 35, so many women over 35). Sudman (1966) describes a method of quota sampling for national face-to-face interview surveys that he termed "probability sampling with quotas". He employed the four quota control groups of men under 35, men 35 and older, employed women and unemployed women, with the control groups chosen to give appropriate representation to young men and employed women. See also Stephenson (1979). The interviewing field force would generally be distributed across the country in a balanced way, either in areas selected to be representative, along the lines employed by Kiær, or in areas selected by a probability sample design. Sometimes additional controls are imposed, for example specifying the routes the interviewers were to follow, with no more than one person sampled in any household. Quota controls can also be applied in telephone surveys, mall intercept surveys, internet surveys (see Section 6), and other types of survey.

Quota sampling has two main advantages over probability sampling: cost and timeliness. Quota sampling is less costly because interviewers do not need to chase up elusive sampled units and because it avoids the costs of sampling specific households or persons (often including the associated listing costs). For the same reasons, a quota sample can be speedily fielded, and the data collected more rapidly than with a probability sample.

Quota sampling is a form of nonprobability sampling that assumes that the respondents in a quota group are an equal probability sample of the population in that group. Note that this assumption also assumes that nonrespondents in the group are missing at random; nonresponse occurs with quota sampling, in essence with respondents substituted for the nonrespondents. Studies that have been conducted to evaluate quota sampling have found that the results are often similar to those produced



by probability sampling, but this is not always the case (see Moser and Stuart, 1953, also Moser and Kalton, 1971; Stephan and McCarthy, 1958). For further references on quota sampling, see Kruskal and Mosteller (1980).

*Random Route Sampling.* Random route, or random walk, sampling is another quasi-probability sampling method that avoids the cost of, and associated time involved with, the listing operation. There are various versions of this method, but each starts with a random selection of a starting household and the interviewers then follow specified rules for walking patterns to follow and selection methods to use for serially identifying the subsequent households. The method has often been used in Europe and it is used in the Expanded Programme of Immunization (EPI) sampling method described in Section 5. Bauer (2014, 2016) discusses the selection errors that can occur with random route sampling and demonstrates that the method does not produce an equal probability sample, as its users generally assume.

## 5. Pseudo-Probability Sample Designs for “Hard-to-Survey Populations”

Recent years have seen a major increase in the use of social survey methods to study the characteristics of “hard-to-survey populations” (Tourangeau, Edwards, Johnson, Wolter, Bates, 2014). Such populations are of various types, but all comprise only a small proportion of the general population and a population for which there is no separate sampling frame. This section presents three examples of sample designs for such populations. The first example is an inexpensive method that has been very widely used for vaccination surveys of the extremely rare population of 1-year-old children. The other two examples describe methods for sampling rare populations where membership of that population is a sensitive characteristic.

### a. *The EPI sampling method.*

For almost 50 years, the World Health Organization’s Expanded Programme on Immunization (EPI) has used simple, inexpensive, sample designs in developing countries for measuring childhood immunization at the district level. Many thousands of EPI surveys have been conducted over this period, and the sample design has evolved over time. The sample design is a two-stage sample of clusters of communities (e.g., villages, towns, health service districts) that are sampled with outdated measures of estimated population sizes, with samples of eligible children selected within selected communities. The standard overall sample size is small, with the selection of 30 clusters and 7 children in each cluster. The design is often known as  $30 \times 7$  design. Except in smaller communities, no household listings are made. Instead, the interviewer goes to the center of the village, chooses a random direction by spinning a bottle on the ground, and counts the number of households in that direction to the edge of the

community. The interviewer then chooses a random number (for instance, from the numbers on a banknote) to identify the first sampled household. The second sampled household is then the one closest to the first, and so on, sequentially until survey data are collected on seven eligible children. Levy and Lemeshow (2008, pp. 427–428) describe the EPI sampling methods and Bennett (1993) describes some of the modifications to the original method.

The US Centers for Disease Control and Prevention (CDC) recommends a probability  $30 \times 7$  sample design for its rapid needs assessment tool, the Community Assessment for Public Health Emergency Response (CASPER) program. In this case, the clusters are generally census blocks with counts of households obtained from the U.S. Census Bureau or by using a GIS program for use in the PPES selection of thirty clusters. The fieldworker counts or estimates the number of households in a sampled cluster, divides that number by seven to give the sampling interval for systematic sampling, proceeds to select the sample from a random starting point, selecting subsequent households using a serpentine walking procedure. A crude weighting adjustment is proposed for use in the data analysis. Details are provided by CDC (2019).

#### *b. Venue-Based Sampling*

Venue-based sampling (also known as location sampling, time-space sampling, center sampling, and intercept sampling) is used for sampling members of a rare population at places that they frequent. It is applicable for rare populations that visit certain locations. It can be used to survey nomadic populations and for sampling hidden rare populations where the membership of that population is a sensitive matter. The method requires the construction of a frame of locations and a decision on the overall time period for the survey, selecting a sample of location/time periods for data collection, and selecting all or a sample of members of the survey population visiting each sampled location in the sampled data collection time period (Kalton, 1991). Two issues of concern arise when sampling hidden populations. One relates to the population coverage provided by the frame of locations and the overall time period: What proportion of the population will fail to visit any of the locations in that time period? Another issue relates to the multiplicity problem: How to account for the variability in the numbers of visits made to any of the locations by different sample members during the overall time period? These numbers are needed for use in weighting to compensate for unequal selection probabilities, but they are unknown. At best, they can be estimated by asking respondents questions about their general frequencies of visiting the locations. See MacKellar, Gallagher, Findlayson, Lansky, and Sullivan (2007) for a description of the sampling methods used for surveying men who have sex with men (MSM) in a number of metropolitan areas in the United States.

### c. Respondent Driven Sampling

Respondent driven sampling (RDS) is a form of link-trace sampling that selects the sample based on the social networks that exist for some populations. RDS has become a popular method for sampling rare hidden populations that have this feature, such as injection drug users and sex workers. The method starts by identifying a small set of members of the population of interest, who serve as *seeds* for the subsequent sample. The seeds respond to the survey, including responding to a question asking how many members of the survey population they know. They are then asked to recruit a set number of members of that population for the survey, the *alters*. The alters then go through the same process, recruiting further sample members. Under idealized circumstances, Heckathorn (1997) has shown that RDS produces a probability sample. However, the many conditions required for this to apply will not hold in practice (Gile and Hancock, 2010).

## 6. Internet Surveys

Recruiting the sample via the internet is a relatively recent approach for conducting social research. This approach has become extremely popular and has led to several alternative methods. See, for example, Baker, Blumberg, Brick *et al.* (2010) for a review of these methods. Surveys based on internet sampling have the great attractions of obtaining responses from large samples at low cost and high speed. However, their nonprobability sampling methods raise concerns about potential biases in the survey estimates. Those without, or with limited, access to the internet are excluded from these surveys and the survey respondents are clearly not a representative sample of the general population.

One form of internet sampling, known as river sampling, attaches invitations to participate in a survey on a number of internet sites, usually with offers of some form of compensation. The biases in the sample selection process make the representativeness of the sample highly questionable. Questions also need to be raised about the honesty and thoughtfulness of the responses.

Another form of internet sampling employs an opt-in internet panel. (An opt-in internet panel is distinct from an internet panel that selects a household panel by probability sampling and then conducts many data collections from the panel over time, albeit typically with low response rates). Extremely large numbers of people are recruited for opt-in internet panels to be available to be approached to respond to surveys over time, sometimes as one of a range of services they may be asked to provide, in exchange for a payment for their services. The panel members can then be selected for invitation to respond to a given survey based on their responses to the screening instrument used in their recruitment.

In some ways, these large-scale nonprobability internet surveys bring to mind the abysmal results obtained from the 1936 Literacy Digest Poll referred to early. However, there are two major differences from the uncontrolled sample in the Digest Poll. One is the attempt to select a representative quota sample in design with internet panels. The other is the use of weighting adjustments in the analysis to achieve the same purpose. Before around 1970, lacking today's computers, complex calibration weighting adjustments were infeasible, but now advanced adjustment methods have been developed and are readily employed for both probability samples (particularly those with low response rates) and for nonprobability samples. With river sampling, a limited number of variables can be collected as part of the data collection for use in calibrating the sample to known or estimated population characteristics. The data collected in the screening instrument for an on-line panel can provide a much greater range of variables that can be used in sample selection and in the application of complex calibration adjustments to make the weighted sample correspond to a wide range of external controls. Nevertheless, serious doubts will persist about whether external data are available for the key auxiliary calibration variables at the population level or for a probability sample of that population, and whether the responses to the on-line survey can be treated as equal to the responses from the external source. Thus, for any given survey estimate, there must be concerns about how representative the nonprobability sample members are of the general population within the controls imposed in design or weighting. There will inevitably remain some residual biases of unknown magnitude and, with large samples, these biases can have a dominant influence on the level of accuracy of the survey estimates (Meng, 2018; Kalton, 2021, pp. 136–137).

## **7. Model-Dependent Inference**

In 1976, Fred Smith—my late friend and colleague at the University of Southampton at that time—wrote a paper reviewing the foundations of survey sampling in which he raised the question of why finite population inference should be so different from inference in the rest of statistics. His view at the time was that 'survey statisticians should accept their responsibility for providing stochastic models for finite populations in the same way as statisticians in the experimental sciences' (Smith, 1976); he moderated his position in a subsequent paper (Smith, 1994). Smith (1976) and papers by Brewer (1963), Royall (e.g., 1970, 1976) and others led to a spirited and longstanding debate about the choice between design-based (model-assisted) inference and model-dependent (or model-based) inference. I was a discussant of Fred's 1976 paper and I subsequently published two papers on the role of models in survey sampling inference, with a defense of design-based inference in most circumstances applicable in large-scale social surveys (Kalton, 1983, 2002). However, models are needed to deal

with the sampling imperfections of noncoverage and nonresponse, and they are needed for subgroup analyses in which the sample sizes are not adequate to provide design-based estimators of adequate precision. With the large decline in response rates that has occurred since the 1970's, it is no longer possible for survey statisticians to treat nonresponse as a minor blemish that can be brushed under the carpet in using design-based inference. I will return to this point later.

The model-dependent approach has led to the development of the prediction approach to survey inference. With this approach, an estimate of the population total  $Y$  is given by

$$\hat{Y}_m = \sum_{i \in S} y_i + \sum_{i \notin S} \hat{f}(X_i) \quad (2)$$

where the first summation is over the observed values in the sample  $S$  of size  $n$  and the second summation is over the model predictions of the  $y$  values for the nonsampled elements in the population. For comparison with the model-assisted design-based estimator  $\hat{Y}_d$  in (1), the model-dependent estimator may be expressed as  $\hat{Y}_m = \sum_S e_i + \sum_U \hat{f}(X_i)$ . In practice, greater care is used to develop the model for  $\hat{Y}_m$  than is the case in developing the working model for  $\hat{Y}_d$ . If the same model is used,  $\hat{Y}_m$  likely has lower variance than  $\hat{Y}_d$ . However,  $\hat{Y}_m$  has a design bias if the model is mis-specified, as is always the case to some extent, and the magnitude of the bias is unknown. The texts by Valliant, Dorfman, and Royall (2000) and Chambers and Clark (2012) describe the prediction approach in detail. The first chapter of Valliant et al. (2000) provides a useful review of design-based and model-based inference and includes further references. Note that the equation for  $\hat{Y}_m$  does not include selection probabilities (except possibly for estimating the model parameters) and does not require a probability sample. However, as Valliant, Dorfman, and Royall (2000, pp. 19–22) argue, randomization has the benefit of giving some protection against imbalance in factors uncontrolled in the design.

In my experience, until recently the prediction approach has had limited utility for large-scale social surveys of households and persons for the following reasons:

1. As distinct from surveys of establishments, there are generally little, if any, data available from the sampling frame about every member of the target population for use in the prediction models. Although some countries maintain up-to-date population registers that contain a selection of individual characteristics, in many countries area sampling is used, with frame construction for individuals or households being performed only in selected areas. In these latter countries, no frame data is available for all members of the target population.
2. Social surveys are multipurpose in nature. They collect survey data on many variables, often numbering in the hundreds, and these data are analyzed in many ways, producing thousands of estimates. As a rule, these surveys are

primarily conducted to produce descriptive estimates of parameters of the survey's finite population. These estimates need to be produced rapidly and to be consistent with each other. (These days, analytic estimates are also often produced, mostly through secondary analyses—see section 7).

3. A large proportion of the variables collected in social surveys are categorical in nature. They often cannot be as well predicted from auxiliary data as is the case with some of the continuous variables collected in business surveys.

However, even with large-scale social surveys, model-dependent estimation has a role to play in the production of descriptive estimates for small subclasses for which the sample sizes are too small to yield design-based estimates of adequate precision. This situation occurs particularly when the subclasses are geographical-defined administrative areas. The growth of interest by policy makers and others in separate estimates for administrative districts of all sizes has led to the development of the subject known as *small area estimation*. For many years, small area estimates, which are obtained using model-dependent prediction methods, were viewed with considerable skepticism by design-based statisticians but they have now become widely accepted in many fields of application. Ghosh (2020) gives a history of the development of small area estimation over five decades and Rao and Molina (2015) give a detailed description of this large and growing field.

The theoretical developments in model-based inference have now become increasingly relevant for social surveys to address the sampling imperfections and limitations with probability samples, and for the analyses of nonprobability samples; the use of nonprobability sampling for social research has grown rapidly in recent years, in particular for internet surveys.

## 8. Analytic Uses of Survey Data

As computing power and software came into widespread use in the 1970's, survey data collected using complex sample designs were used, mostly in secondary analyses, to produce analytic statistics that studied the relationships between variables, often looking for causal connections. Initially, multiple regression was the main form of analysis, with interest directed to the magnitude of the regression coefficients. Many analysts argued that their interest in the results of these analyses was not for the specific finite population surveyed, but rather as estimates of superpopulation parameters of universal generality, and that, with the "correct" model, aspects of the sample design were irrelevant. From this perspective, probability sampling of the finite population becomes irrelevant and, unless survey weights and clustering were important as predictor variables, their inclusion in the analysis in a standard design-based way serves only to lower the precision of the estimated regression coefficients. The counter

position was that no model is totally correct and that the estimation of the population regression coefficients, often termed census parameters, using the survey weights provides a safer approach. There is extensive literature on this topic. See, for example, DuMouchel and Duncan (1983).

Over time, the use of regression methods with survey data has been extended to include a wide range of regression models and other multivariate analysis techniques such as categorical data analysis, multilevel modeling, and longitudinal analyses. It is outside the scope of this paper to describe the application of these methods with complex survey data. See Skinner, Holt, and Smith (1989), Chambers and Skinner (2003). Applications of a range of multivariate methods with complex survey data are well described in the texts by Korn and Graubard (1999) and Heeringa, West, and Berglund (2017).

## **9. Administrative Records and Big Data**

A great deal of attention has been paid recently to the use of administrative records as an alternative source of research data. There are obvious serious issues of privacy and confidentiality to be addressed when government-maintained administrative data are used in this way. For this reason, this approach is particularly suited to researchers in government agencies. The approach has notable potential attractions in terms of cost and sample size, but it needs to be recognized that it has its limitations. For instance, what is the coverage of the frame of the records, especially regarding program enrollment versus eligibility? Do the records contain the data needed to measure the concepts as the researcher would like to define them? Are the record data measured consistently across the population, or are there differences in the procedures used in different administrative areas? Are the data measured consistently over time to enable time series data to be validly analyzed? How might changes in program rules affect temporal comparisons? How long is the period between data collection and the researcher's access to an analyzable dataset? Do the records contain the full set of variables needed for the analyses? In many cases, a single set of administrative records does not contain all the variables needed for the analyses. In this situation, it may be possible to link two or more sets of records, but record linkage problems need to be overcome and greater issues of confidentiality must be addressed.

How accurate are the data recorded in the records? Survey researchers have devoted a great deal of effort to training a relatively small number of interviewers to ask and record respondents' answers in a standard way. The situation is different with administrative records. Charlie Cannell, my late friend and colleague at the University of Michigan's Survey Research Center, had the following quotation from Josiah Stamp (1880–1941) in a plaque on his office wall:

“The government are very keen on amassing statistics. They collect them, add them, raise them to the  $n$ th power, take the cube root and prepare wonderful diagrams. But you must never forget that every one of these figures comes in the first instance from the village watchman, who just puts down what he damn pleases.”

While not claiming that current administrative records are as bad as this quotation might suggest, those who use such records for statistical purposes should carefully assess their quality and the distortions to which they may be subjected. See the paper by Hand (2018) and the ensuing discussion for a detailed discussion of the advantages and limitations of administrative records for research purposes.

In addition to government-maintained administrative records, there are other sources of social research data. In some cases, nongovernment records, such as those maintained by private organizations, may contain relevant information. However, they are subject to similar quality concerns, and access to the records may be hard to obtain. There are also sources of big data that occur on a flow basis, such as from linking cell phones to their GPS locations. The cell phone locations can be used to provide information about commuter times and even about long-distance travel trips if the home location is identified. Another source of big data is from scrapings on the web. Google Flu Trends (GFT) is a well-known and cautionary example. By analyzing extremely large numbers of flu-related searches on the web, Google developed models to predict local flu outbreaks in real time, avoiding the inevitable delay with other data sources. However, the models have since been found to fail (Lazer, Kennedy, King, and Vespignani, 2014), which serves as a warning that the apparent attraction of very big datasets can be illusory. For another example, see Bradley, Kuriwaki, Isakov, Sejdinovic, Meng, and Flaxman (2021).

## **10. Concluding Remarks**

As illustrated in previous sections, the choice between purposive selection and probability sampling was a subject of debate in the early period of survey research. It was not until after Neyman's (1934) paper that probability sampling and design-based inference were established as the gold standard for large-scale surveys conducted by national statistical offices. With a perfectly executed probability sample and no response error, the analyst has the security of being able to report the survey findings as being subject only to a measurable degree of sampling error, whereas with nonprobability sampling the analyst can always be challenged that a purposive sample is not representative of the population with respect to the variables of analytic interest.

The preeminence of probability sampling for government surveys in the years from 1940 to, say, 2010 was not universal. There are costs incurred with probability sampling



and a probability sample takes more time to draw and data collection takes longer. As illustrated in earlier sections, failures to devise probability sampling methods that can be applied with acceptable cost and timeliness for certain populations has given rise to the development of shortcut methods that depart in varying degrees from rigorous probability sampling.

In the early days, the idea of a “representative sample” was restricted to a sample that was representative in its design, as was the case with Kiær’s designs. The use of weighting adjustments in the analysis to achieve representativeness was seldom considered. The failure of the Literacy Digest poll in predicting the result of the U.S. Presidential election made clear that an extremely large unrepresentative sample could, without weighting adjustments, yield bad results.

Over the years, the implementation of probability sampling in social surveys has been increasingly challenged in many—but not all—countries by a steady decline in the willingness of the public to participate in surveys. Despite greater efforts to encourage response, response rates have declined dramatically in recent years. In reaction, greater efforts have been made to compensate for nonresponse, with major advances in the techniques employed. While replication methods of variance estimation can be applied to reflect the effect of the use of these techniques on the precision of the survey estimates, their use results in lower precision. Furthermore, the nonresponse adjustment model cannot be assumed to be “correct,” and the extent of any remaining nonresponse bias cannot be assessed. With its current heavy reliance on nonresponse models, in many countries probability sampling with design-based inference no longer retains its status as the undisputed gold standard. Moreover, the current levels of nonresponse have led to a marked increase in the costs of conducting a survey with probability sampling, both because of the increase in the initial sample size needed to produce the required sample size and because of the increased efforts to counteract nonresponse. For example, in the U.S. random digit dialing (RDD) was widely used with telephone surveying in the later part of the last century and the early part of this one because of the cost-efficiency of this modality (particularly for surveying rare populations). However, response rates for RDD surveys have plummeted to a level as low as 10 to 20 percent, largely ruling out this form of sampling.

With the security of model-free probability sampling with design-based inference now a thing of the past, model-dependent methods appear to be taking on a major role in social statistics. Research on making valid inferences from nonprobability samples is ongoing (see, for example, Valliant, 2020). Models are increasingly used to analyze data from a combination of data sources, including survey data from probability and nonprobability samples, administrative records, and other sources of big data. Thus, there is much research currently underway on making inferences from combinations

of probability and nonprobability samples and from probability samples and other data sources (Kim and Wang, 2019; Beaumont and Rao, 2021; Rao, 2021),

In summary, after a long period in which probability sampling methods have dominated, the current situation is in a state of flux. New methods involving nonprobability sampling, internet sampling, administrative records, and big data are under constant modification and development. Brackstone (1999) lists six aspects of data quality for a statistical agency that remain applicable: relevance (how well the data meet the needs of the clients); accuracy (including both bias and variance); timeliness (time between the reference point and the time of data availability); interpretability (availability of relevant metadata); and coherence (ability to bring the data into a broader framework, including over time). The new data collection methods need to be assessed against these measures and, furthermore, the extensive research on response errors that has been conducted in the past now needs to be applied with the new methods of data collection. This is an exciting and challenging time for survey methodologists.

## References

- Aldrich, J., (2008). Professor A. L. Bowley's theory of the representative method. (Discussion Papers in Economics and Econometrics, 801) University of Southampton. <https://eprints.soton.ac.uk/150493>.
- Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D., Frankel, M. R., Garland, G., Groves, R. M., Kennedy, C., Krosnick, J., Lavrakas, P. J., (2010). AAPOR Report on Online Panels. *Public Opinion Quarterly*, 74(4), pp. 711–781.
- Bauer J. J., (2014). Selection errors of random route samples. *Sociological Methods and Research*, 43(3), pp. 519–544.
- Bauer J. J., (2016). Biases in random route surveys. *Journal of Survey Statistics and Methodology*, 4(2), pp. 263–287.
- Beaumont J-F., Rao, J. N. K., (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? *The Survey Statistician*, 83, pp. 11–22.
- Bennett, S., (1993). Cluster sampling to assess immunization: a critical appraisal. *Bulletin of the International Statistical Institute*, 49<sup>th</sup> Session, 55(2), pp. 21–35.
- Bowley, A. L., (1913). Working-class households in Reading. *Journal of the Royal Statistical Society*, 76(7), pp. 672–701.

- Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L., Flaxman, S., (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600, pp. 695–700.
- Brewer, K. R. W., (1963). Ratio estimation in finite populations: some results deducible from the assumption of an underlying stochastic process, *Australian Journal of Statistics*, 5, pp. 93–105.
- Caradog Jones, D., (1949). *Social Surveys*. Hutchinson's University Library, London.
- CDC, (2019). Community Assessment for Public Health Emergency Response (CASPER) Toolkit. 3<sup>rd</sup> ed., CDC, Atlanta. <https://www.cdc.gov/nceh/casper/>.
- Chambers, R., Clark, R., (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford University Press, Oxford.
- Chambers, R. L., Skinner, C. J., Eds., (2003). *Analysis of Survey Data*. Wiley, Chichester.
- Cochran, W. G., (1953). *Sampling Techniques*. Wiley, New York.
- Converse, J. M., (2017). *Survey Research in the United States: Roots and Emergence 1890-1960*. Routledge, New York.
- DuMouchel, W. H., Duncan, G. J., (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, 78, pp. 535–543.
- Ghosh, M., (2020). Small area estimation: its evolution in five decades (with discussion). *Statistics in Transition*, 21(4), pp. 1–67.
- Gile, K. J., Hancock, M. S., (2010). Respondent-driven sampling: an assessment of current methodology. *Sociological Methodology*, 40(1), pp. 285–327.
- Gini, C., Galvani, L., (1929). Di una applicazione del metodo representative. *Annali di Statistica*, 6(4), pp. 1–107.
- Hand, D. J., (2018). Statistical challenges of administrative and transaction data (with discussion). *Journal of the Royal Statistical Society, A*, 181(3), pp. 555–605.
- Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953). *Sample Survey Methods and Theory. Volume I: Methods and Applications. Volume II: Theory*. Wiley, New York.
- Heckathorn, D. D., (1997). Respondent-driven sampling: a new approach to the study of hidden populations. *Social Problems*, 44(2), pp. 174–199.
- Heeringa, S. G., West, B. T., Berglund, P. A., (2017). *Applied Survey Data Analysis*. Chapman & Hall/ CRC, Boca Raton, FL.

- Jensen, A., (1926) The report on the representative method in statistics. *Bulletin of the International Statistical Institute*, 22, pp. 355–376.
- Kalton, G., (1983). Models in the practice of survey sampling. *International Statistical Review*, 51, pp. 175–188.
- Kalton, G., (1991). Sampling flows of mobile human populations. *Survey Methodology*, 17(2), pp. 183–194.
- Kalton, G., (2002). Models in the practice of survey sampling (revisited). *Journal of Official Statistics*, 18, pp. 129–154.
- Kalton, G., (2019). Developments in survey research over the past 60 years: A personal perspective. *International Statistical Review*, 87 (S1), pp. S10–S30.
- Kalton, G., (2021). *Introduction to Survey Sampling*. 2<sup>nd</sup> ed. SAGE Publications, Thousand Oaks, California.
- Kiær, A. N., (1976). *The Representative Method of Statistical Surveys*. English translation, Statistisk Centralbyro, Oslo.
- Kim, J. K., Wang, Z., (2019). Sampling techniques for big data analysis. *International Statistical Review*, 87 (S1), pp. S177–S191.
- Kish, L., (1995). The hundred years' war of survey sampling. *Statistics in Transition*, 2(5), pp. 813–830.
- Korn, E. L., Graubard, B. I., (1999). *Analysis of Health Surveys*. Wiley, New York.
- Kruskal, W., Mosteller, F., (1980). Representative sampling, IV: The history of the concept in statistics, 1895-1939. *International Statistical Review*, 48(2), pp. 169–195.
- Lazer, D., Kennedy, R., King, G., Vespignani, A., (2014). The parable of Google flu: Traps in big data analysis. *Science*, 343, pp. 1203–1205.
- Levy, P. S., Lemeshow, S., (2008). *Sampling of Populations. Methods and Applications*. 4<sup>th</sup> ed. Wiley, Hoboken, NJ.
- Lie, E., (2002). The rise and fall of sampling surveys in Norway, 1875–1906. *Science in Context*, 15(3), pp. 385–409.
- Lohr, S. L., Brick, J. M., (2017). Roosevelt predicted to win: Revisiting the 1936 Literary Digest Poll. *Statistics, Politics, and Policy*, 8(1), pp. 65–84.
- MacKellar, D. A., Gallagher, K. M., Findlayson, T., Sanchez, T., Lansky, A., Sullivan, P. S., (2007). Surveillance of HIV risk and prevention behaviors of men who have sex with men—a national application of venue-based, time-space sampling. *Public Health Reports*, 122 (1), Supplement 1, pp. 39–47.

- Mahalanobis, P. C., (1946). Recent experiments in statistical sampling in the Indian Statistical Institute (with discussion). *Journal of the Royal Statistical Society*, 109, pp. 325–378.
- Meng, X-L., (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics*, 12(2), pp. 685–726.
- Moser, C. A., Kalton, G., (1971). *Surveys Methods in Social Investigation*. 2<sup>nd</sup> ed. Heinemann, London.
- Moser, C. A., Stuart, A., (1953). An experimental study of quota sampling. *Journal of the Royal Statistical Society*, A, 116, pp. 349–405.
- Neyman, J., (1934). On two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, pp. 558–625.
- Rao, J. N. K., (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, 83, pp. 242–272.
- Rao, J. N. K., Molina, I., (2015). *Small Area Estimation*. 2<sup>nd</sup> ed. Wiley, Hoboken, N. J.
- Royall, R. M., (1970). On finite population sampling theory under certain regression models. *Biometrika*, 57, pp. 377–387.
- Royall, R. M., (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, pp. 657–664.
- Särndal, C.E., Swensson, B., Wretman, J., (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.
- Skinner, C. J., Holt, D., Smith, T. M. F., Eds., (1989). *Analysis of Complex Surveys*. Wiley, Chichester.
- Smith, T. M. F., (1976). The foundations of survey sampling: a review. *Journal of the Royal Statistical Society*, A, 139, pp. 183–204.
- Smith, T. M. F., (1994). Sample surveys 1975-90; an age of reconciliation? *International Statistical Review*, 62, pp. 5–34.
- Stephan, F. F., (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association*, 43(241), pp. 12–39.
- Stephan, F. F., McCarthy P. J., (1958). *Sampling Opinions. An Analysis of Survey Procedures*. Wiley, New

- Stephenson, C. B., (1979). Probability sampling with quotas: An experiment. *Public Opinion Quarterly*, 43(4), pp. 477–497.
- Sudman, S., (1966). Probability sampling with quotas. *Journal of the American Statistical Association*, 61, pp. 749–771.
- Tourangeau, R., Edwards, B., Johnson, T. P., Wolter, K.M., Bates, N., Eds., (2014). *Hard-to-Survey Populations*. Cambridge University Press, Cambridge, U. K.
- Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2), pp. 231–263.
- Valliant, R., Dorfman, A. H., Royall, R. M., (2000). *Finite Population Sampling and Inference. A Prediction Approach*. Wiley, New York.
- Yates, F., (1949). *Sampling Methods for Censuses and Surveys*. Griffen, London.

## **Comments on „Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day” by Graham Kalton**

**Danny Pfeffermann<sup>1</sup>**

I like to congratulate Professor Kalton for writing this very constructive article on probability versus nonprobability sampling. I learned a lot from reading it. In what follows, I add a few comments on this topic.

1- Professor Kalton emphasizes the issue of representative samples. In my view, probability samples and obviously nonprobability samples are practically never representative, even if balanced in advance on certain control (covariate) variables. A major reason for this is nonresponse, which might be “not missing at random” (NMAR), in which case the response probabilities depend on the target study variable, even after conditioning on known covariates. However, even in the case of simple random sampling and complete response, the actual sample may not be representative with respect to the unknown study variables, simply because of the randomness of the sample selection, unless the sample size is sufficiently large. Clearly, this problem worsens when sampling with unequal probabilities. Classical design-based theory overcomes this problem by restricting the inference to the randomization distribution over all possible sample selections. Thus, an estimator of a population mean is unbiased if its average over all possible samples that could have been drawn equals the true population mean, but in practice, we only have one sample. The use of models does not solve this problem either. A good model has to account for the sampling probabilities and the model assumed for the population values, and the inference need to account for both stochastic processes. As illustrated in many articles, ignoring the sampling process when fitting models to the sample data results with biased estimators of the model parameters in the case of informative sampling, by which the sampling probabilities are correlated with the outcome variables, again after conditioning on the model covariates. See, e.g. Pfeffermann and Sverchkov (1999) for empirical illustrations. In the case of NMAR nonresponse, the model has to account also for the unknown response probabilities.

---

<sup>1</sup> Department of Statistics, Hebrew University, Jerusalem, Israel & Southampton Statistical Sciences Research Institute, University of Southampton, UK. E-mail: [msdanny@mail.huji.ac.il](mailto:msdanny@mail.huji.ac.il); [msdanny@soton.ac.uk](mailto:msdanny@soton.ac.uk).  
ORCID: <https://orcid.org/0000-0001-7573-2829>.



2- The problem of nonresponse is indeed troubling and requires the use of models in the case of NMAR nonresponse, even in the case of design-based inference. The use of a response model enables to adjust the base sampling weights by the inverse of the estimated response probabilities, viewed as a second stage of the sampling process. I should say though that unlike a common perception, the response model can be tested, by testing the model of the study variable holding for the responding units, which accounts for the sampling design and the response. See, e.g. Pfeffermann and Sikov (2011).

3- Professor Kalton discusses the pros and cons of internet surveys “standing on their own”. I like to add that internet surveys are often used as one, out of several possible modes of response. For example, a questionnaire is sent to all the sampled units. It encourages them to respond via the internet. Those who do not respond are approached by telephone. When no response is obtained, an interviewer is sent for a face-to-face interview.

A well-known problem with this procedure is of mode effects; different estimates obtained from the respondents to the different modes, either because of differences between the characteristics of respondents responding with the different modes, (selection effect), or because of responding differently by the same sampled unit, depending on the mode of response (measurement effect). Several approaches to deal with this problem have been proposed in the literature. See, e.g. De Leeuw et al. (2018) for a comprehensive review.

My last 2 comments refer to inference from nonprobability samples:

4- Denote by  $S_{NP}$  the nonprobability sample. Rivers (2007) proposes to deal with the possible non-representativeness of  $S_{NP}$  by the use of sample matching. (Rivers considers a Web sample as the nonprobability sample but here I extend the idea to a more general nonprobability sample.) The approach consists of using a probability (reference) sample  $S_R$  from the target population, drawn with probabilities  $\pi_k = \Pr(k \in S_R)$ , and matching to every unit  $i \in S_R$  an element  $k \in S_{NP}$ , based on known auxiliary (matching) variables  $x$ . Denote by  $S_M$  the matched sample. Suppose that it is desired to estimate a population total of a study variable  $Y$ , based on measurements  $\{\tilde{y}_j, j \in S_{NP}\}$ . Estimate,  $\hat{Y}_T = \sum_{j \in S_M} w_j \tilde{y}_j$ ;  $w_j = (1/\pi_j)$ . Clearly, the base sampling weights can be modified to account for nonresponse.

This is an intriguing approach, but its success depends on the existence of a reference probability sample  $S_R$ , which allows sufficiently close matching, and ignorability of membership in the nonprobability sample  $S_{NP}$ , conditional upon



the matching variables. I do not know whether this approach is used in practice, but I think that it deserves further investigation, with proper modifications.

- 5- The last two decades have witnessed the rapid growing of data science. One of the facets of this growth is that some people are agitating that the existence of all sorts of “big data” and the new advanced technologies that have been developed to handle these data, will soon replace the use of sample surveys. In an article I published in 2015, I overviewed some of the problems with the use of big data for the production of official statistics but clearly, when such data sources are available, accessible and timely, they cannot and should not be ignored. Big data can be viewed as a big, nonprobability sample, which for all kinds of reasons is not representative of the target population, and relying just on them can yield biased inference. Integrating big data with surveys is a major issue for research. See, e.g. Kim and Zhonglei (2018) and Rao (2021) for possible approaches, with references to other studies.

***I conclude my discussion by congratulating Statistics in Transition for its 30<sup>th</sup> anniversary and the publication of its 100<sup>th</sup> issue. This is one of the best journals of its kind and I wish it to continue prospering in the coming years.***

## References

- De Leeuw, E. D., Suzer-Gurtekin, Z. and Hox, J., (2018). The Design and Implementation of Mixed Mode Surveys. In *Advances in Comparative Survey Methodology*. Wiley, New York.
- Kim, J. K. and Zhonglei Wang, (2008). Sampling Techniques for Big Data. *International Statistical Review*, 87, pp. 177–191.
- Pfeffermann, D., (2015). Methodological Issues and Challenges in the Production of Official Statistics. *The Journal of Survey Statistics and Methodology (JSSAM)*, 3, pp. 425–483.
- Pfeffermann, D. and Sverckov, M., (1999). Parametric and Semi-Parametric Estimation of Regression Models Fitted to Survey Data. *Sankhya*, 61, pp. 166–186.
- Pfeffermann, D. and Sikov, A., (2011). Imputation and Estimation under Nonignorable Nonresponse in Household Surveys with Missing Covariate Information. *Journal of Official Statistics*, 27, pp. 181–209.
- Rao, J. N. K., (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, 83, pp. 242–272.
- Rivers, D., (2007). Sampling for Web Surveys. Joint Statistical Meetings, Proceedings of the Section on Survey Research Methods.



## **Comments on „Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day” by Graham Kalton**

**Risto Lehtonen<sup>1</sup>**

I would like to congratulate Professor Graham Kalton for his significant and inspiring article entitled as "Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day". The article provides an elegant overview of the history of survey sampling, covering the purposive approaches that dominated the sampling field in the early days but from the 1940s, at least in official statistics, were gradually replaced entirely by probability-based approaches. Today we may be facing a paradigm shift again, but the direction is the opposite. Non-probability-based approaches are becoming viable, if not the only option, in fields that are moving towards big data and other new data sources and new methodological approaches.

The country's data infrastructure forms the basis of official statistics and opens up for me an important perspective on Kalton's presentation. Both probability and non-probability sampling and inference can benefit from statistical data infrastructures that contain a rich selection of micro-level covariates drawn from a variety of administrative and other registers. Perhaps the best options are in countries where population data from register sources and sample data are linked for combined micro-level databases. However, the utility of model-based (prediction) approaches for large-scale social surveys of households and persons will be limited if unit-level data for population members is missing from the sampling frames, as pointed out by Prof. Kalton. This is an important point and I think it can be extended to design-based model-assisted approaches that use mixed models in particular.

Countries differ much in terms of infrastructures based on administrative data. For example, Constance Citro calls for a move to multiple data sources that include administrative records and, increasingly, transaction and Internet-based data (Citro 2014). Eric Rancourt argues that Statistics Canada is facing the new data world by modernizing itself and embracing an admin-first (in the broadest sense) paradigm as a statistical paradigm for the agency (Rancourt 2018). According to the United

---

<sup>1</sup> University of Helsinki, Finland. E-mail: [risto.lehtonen@helsinki.fi](mailto:risto.lehtonen@helsinki.fi).



Nations Economic Commission for Europe (UNECE) report on register-based statistics in the Nordic countries, Central Population Registers of Denmark, Finland, Norway and Sweden were established in the sixties, and for example a totally register-based census was first implemented in Denmark (1981) and next in Finland (1990) (UNECE 2007). The number of national statistical institutes that have adopted or are developing administrative data infrastructures is increasing, as also described in the UNECE report on the use of registers and administrative data for population and housing censuses (UNECE 2018). This development can enhance the use of methods that utilize modeling and individual-level population frame data for model-assisted or prediction-based estimation with probability-based or non-probability-based sample data sets and their combinations.

The situation is different in countries that do not have similar high-quality population registers as for example in the Nordic countries. A recent contribution by Dunne and Zhang (2023) provides one important methodological approach for such countries. The authors present an innovative system (the PECADO application) for population estimates compiled from administrative data only.

Today, in the Nordic countries, as Finland, a majority of official statistics are based on administrative register combinations. In Finland, official statistics are produced by 13 expert organisations in the field of public administration and is coordinated by Statistics Finland. Probability samples are mainly used for regular social surveys such as labour force surveys and special surveys, e.g. Time Use survey. In these surveys, the sample elements can be uniquely linked with the elements in the register databases that often contain a lot of important background data including demographic, regional, socio-economic, income, educational, labour force status, and other variables. Thus these data need not to be collected by direct data collection methods from the respondents, and measurement errors are avoided. In addition, these variables are also used for calibration and model-assisted estimation procedures.

As an example, let me describe briefly the sampling and estimation design of the Labour Force Survey (LFS) of Finland. According to the quality description, in most European countries the LFS is based on a sample of households, and all members of a sample household living at the same address are interviewed. Finland is one of the Nordic countries where LFS is based on sampling of individual persons. The sample of about 12,500 persons is drawn by stratified probability sampling from Statistics Finland's population database, which is based on the Central Population Register. Auxiliary information from registers include gender, age, region and language and selected register variables on employment, completed education and degrees, and income from the Employment Service Statistics of the Ministry of Economic Affairs and Employment, Statistics Finland's Register of Completed Education and Degrees, and the Tax Administration's Incomes Register (Quality Description: Labour Force

Survey, Statistics Finland 2022). Sample data are linked to data from the registry using unique ID keys that exist across all data sources and are used in estimation procedures, including nonresponse adjustments. My experience is that this type of data infrastructure can also provide an excellent sampling and auxiliary data platform for e.g. methodological research in survey statistics; see for example Lehtonen, Särndal and Veijanen (2003, 2005).

Data infrastructures based on integrated administrative and other registers should be based on appropriate statistical theory and methodology for quality assessment and control and quality improvement. Recent sources in the field are for example Zhang (2012), Zhang and Haraldsen (2022) and the book on register-based statistics by Anders Wallgren and Britt Wallgren (2014). Research in statistical data integration and data science methods relevant for official statistics also is extending. A recent source is Yang and Kim (2020).

Experiences show that data infrastructures for official statistic containing a wealth of micro-level information on the population and an option for integration of the various register and sample data sources provide a flexible and efficient framework for survey estimation with probability-based samples. For non-probability samples, the variables of interest are typically in the non-probability data source. Most current methods for valid inference require an auxiliary data source containing the same covariates as the non-probability sample. These data can be obtained from the statistical population register or, more commonly, from a probability sample from it (e.g. Kim, Park, Chen and Wu 2021; Wu 2022). It can be foreseen that although the golden age of probability sampling may be over, probability sampling and non-probability sampling are not in conflict, but can complement each other.

## References

- Citro, C. F., (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, 40(2), pp. 137–161.
- Dunne, J. and Zhang, L.-C., (2023). A system of population estimates compiled from administrative data only. *Journal of the Royal Statistical Society Series A: Statistics in Society*. <https://doi.org/10.1093/jrssa/qnad065>.
- Kim, J.-K., Park, S., Chen, Y. and Wu, C., (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184, pp. 941–963.
- Lehtonen, R., Särndal, C.-E. and Veijanen, A., (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, 29(1), pp. 33–44.

- Lehtonen, R., Särndal, C.-E. and Veijanen, A., (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, 7(3), pp. 649–673.
- Quality Description: Labour force survey, Statistics Finland 2022, (2022). [https://www.tilastokeskus.fi/til/tyti/2022/01/tyti\\_2022\\_01\\_2022-02-22\\_laa\\_001\\_en.html](https://www.tilastokeskus.fi/til/tyti/2022/01/tyti_2022_01_2022-02-22_laa_001_en.html)
- Rancourt, E., (2018). *Admin-First as a statistical paradigm for Canadian official statistics: Meaning, challenges and opportunities*. Proceedings of Statistics Canada Symposium 2018.
- United Nations Economic Commission for Europe, (2007). *Register-based statistics in the Nordic countries: Review of best practices with focus on population and social statistics*. United Nations, New York. <https://digitallibrary.un.org/record/609979?ln=en>
- UNECE, (2018). *Guidelines on the use of registers and administrative data for population and housing censuses*. United Nations, New York and Geneva. <https://unece.org/guidelines-use-registers-and-administrative-data-population-and-housing-censuses-0>
- Yang, S. and Kim, J. K., (2020). Statistical data integration in survey sampling: a review. *Jpn J Stat Data Sci*, 3, pp. 625–650.
- Zhang, L.-C., (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66(1), pp. 41–63.
- Zhang, L.-C. and Haraldsen, G., (2022). Secure big data collection and processing: framework, means and opportunities. *Journal of the Royal Statistical Society: Series A*, *Statistics in Society*, (In Press).
- Wallgren, A. and Wallgren, B., (2014). *Register-Based Statistics: Administrative Data for Statistical Purposes*. Second edition. Wiley.
- Wu, C., (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, 48(2), pp. 283–311.

## **Discussion of “Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day” by Graham Kalton**

**Julie Gershunskaya<sup>1</sup>, Partha Lahiri<sup>2</sup>**

In this excellent overview of the history of probability and nonprobability sampling from the end of the nineteenth century to the present day, Professor Graham Kalton outlines the essence of past endeavors that helped to define philosophical approaches and stimulate the development of survey sampling methodologies. From the beginning, there was an understanding that a sample should, in some ways, resemble the population under study. In Kær’s ideas of “representative sampling” and Neyman’s invention of probability-based approach, the prime concern of survey sampling has been to properly plan for representing characteristics of the finite population. Poststratification and other calibration methods were developed for the same important goal of better representation.

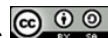
Professor Kalton’s paper underscores growing interest in the use of nonprobability surveys. With recent proliferation of computers and the internet, wealth of data becomes available to researchers. However, “opportunistic” information collected with present-day capabilities usually is not purposely planned or controlled by survey statisticians. No matter how big such a nonprobability sample could be, it may inaccurately reflect the finite population of interest, thus presenting a substantial risk of an estimation bias.

Below, we discuss several recent papers that propose ways to incorporate nonprobability surveys to produce estimates for both large and small areas. Specifically, we will consider two situations often encountered in practice. In the first situation, a nonprobability sample contains the outcome variable of interest, and the main task is to reduce the selection bias with the help of a reference probability sample that does not contain the outcome variable of interest. In the second situation, a probability sample contains the outcome variable of interest, but there is little or no sample available to produce granular level estimates. For such a small area estimation problem, we consider a case when we have access to a large nonprobability sample that does not contain the outcome variable but contains some related auxiliary variables also present in the probability sample. In both situations, researchers have discussed statistical data integration techniques in which a reference probability sample is combined with a nonprobability sample in an effort to overcome deficiencies associated with both probability and nonprobability samples.

---

<sup>1</sup>U.S. Bureau of Labor Statistics, 2 Massachusetts Ave NE Washington, DC 20212, USA, E-mail: [Gershunskaya.Julie@bls.gov](mailto:Gershunskaya.Julie@bls.gov). ORCID: <https://orcid.org/0000-0002-0096-186X>.

<sup>2</sup>University of Maryland, College Park, MD 20742. USA. E-mail: [plahiri@umd.edu](mailto:plahiri@umd.edu). ORCID: <https://orcid.org/0000-0002-7103-545X>.



One way to account for the selection bias of a nonprobability sample is by estimating the sample inclusion probabilities, given available covariates. Then, the inverse values of estimated inclusion probabilities are used, in a similar manner as the usual probability sample selection weights, to obtain estimates of target quantities. Several approaches to estimation of nonprobability sample inclusion probabilities (or propensity scores) have been considered in the literature. Recent papers by Chen et al. (2020), Wang et al. (2021), and Savitsky et al. (2022) propose ways to estimate these probabilities based on combining nonprobability and probability samples. Kim J. and K. Morikawa (2023) propose an empirical likelihood based approach under a different setting. To save space, we will not discuss their approach. We now review three statistical data integration methods.

The approaches concern with the estimation of probabilities  $\pi_{ci}(x_i) = P\{c_i = 1 | x_i\}$  to be included into the nonprobability sample  $S_c$ , for units  $i = 1, \dots, n_c$ , where  $c_i$  is the inclusion indicator of unit  $i$  taking on the value of 1 if unit  $i$  is included into the nonprobability sample, and 0 otherwise;  $x_i$  is a vector of known covariates for unit  $i$ ;  $n_c$  is the total number of units in sample  $S_c$ . The problem, of course, is that we cannot estimate  $\pi_{ci}$  based on the set of units in nonprobability sample  $S_c$  alone, because  $c_i = 1$  for all  $i$  in  $S_c$ . The probabilities are estimated by combining set  $S_c$  with a probability sample  $S_r$ . Due to its role in this approach, the probability sample here is also called “reference sample”.

Assuming both nonprobability and probability samples are selected from the same finite population  $P$ , Chen et al. (2020) write a log-likelihood, over units in  $P$ , for the Bernoulli variable  $c_i$ :

$$\ell_1(\boldsymbol{\theta}) = \sum_{i \in P} \{c_i \log [\pi_{ci}(x_i, \boldsymbol{\theta})] + (1 - c_i) \log [1 - \pi_{ci}(x_i, \boldsymbol{\theta})]\}, \quad (1)$$

where  $\boldsymbol{\theta}$  is the parameter vector in a logistic regression model for  $\pi_{ci}$ .

Since finite population units are not observed, Chen et al. (2020) employ a clever trick and re-group the sum in (1) by presenting it as a sum of two parts: part 1 involves the sum over the nonprobability sample units and part 2 is the sum over the whole finite population:

$$\ell_1(\boldsymbol{\theta}) = \sum_{i \in S_c} \log \left[ \frac{\pi_{ci}(x_i, \boldsymbol{\theta})}{1 - \pi_{ci}(x_i, \boldsymbol{\theta})} \right] + \sum_{i \in P} \log [1 - \pi_{ci}(x_i, \boldsymbol{\theta})]. \quad (2)$$

Units in part 1 of the log-likelihood in (2) are observed; for part 2, Chen et al. (2020) employ the pseudo-likelihood approach by replacing the sum over the finite population with its probability sample based estimate:

$$\hat{\ell}_1(\boldsymbol{\theta}) = \sum_{i \in S_c} \log \left[ \frac{\pi_{ci}(x_i, \boldsymbol{\theta})}{1 - \pi_{ci}(x_i, \boldsymbol{\theta})} \right] + \sum_{i \in S_r} w_{ri} \log [1 - \pi_{ci}(x_i, \boldsymbol{\theta})], \quad (3)$$

where weights  $w_{ri} = 1/\pi_{ri}$  are inverse values of the reference sample inclusion probabilities  $\pi_{ri}$ . Estimates are obtained by solving respective pseudo-likelihood based estimating equations.

One shortcoming of the Chen et al. (2020) approach is that their Bernoulli likelihood is formulated with respect to an unobserved indicator variable. Although the regrouping



employed in (2) helps to find a solution, results obtained by Wang et al. (2021) indicate that it is relatively inefficient, especially when the nonprobability sample size is much larger than the probability sample size.

Wang et al. (2021) formulate their likelihood for an *observed* indicator variable and thus their method is different from the approach of Chen et al. (2020). To elaborate, Wang et al. (2021) introduce an imaginary construct consisting of two parts: they *stack* together non-probability sample  $S_c$  (part 1) and finite population  $P$  (part 2). Since nonprobability sample units belong to the finite population, they appear in the stacked set twice. Let indicator variable  $\delta_i = 1$  if unit  $i$  belongs to part 1, and  $\delta_i = 0$  if  $i$  belongs to part 2 of the stacked set; the probabilities of being in part 1 of the stacked set are denoted by  $\pi_{\delta_i}(x_i) = P\{\delta_i = 1|x_i\}$ . Wang et al. (2021) assume the following Bernoulli likelihood for observed variable  $\delta_i$ :

$$\ell_2(\tilde{\theta}) = \sum_{i \in S_c} \log \left[ \pi_{\delta_i}(x_i, \tilde{\theta}) \right] + \sum_{i \in P} \log \left[ 1 - \pi_{\delta_i}(x_i, \tilde{\theta}) \right], \tag{4}$$

where  $\tilde{\theta}$  is the parameter vector in a logistic regression model for  $\pi_{\delta_i}$ . Since the finite population is not available, they apply the following pseudo-likelihood approach:

$$\hat{\ell}_2(\tilde{\theta}) = \sum_{i \in S_c} \log \left[ \pi_{\delta_i}(x_i, \tilde{\theta}) \right] + \sum_{i \in S_r} w_{ri} \log \left[ 1 - \pi_{\delta_i}(x_i, \tilde{\theta}) \right]. \tag{5}$$

Existing ready-to-use software can be used to obtain estimates of  $\pi_{\delta_i}$ . However, the actual goal is to find probabilities  $\pi_{ci}$  rather than probabilities  $\pi_{\delta_i}$ . Wang et al. (2021) propose a two-step approach, where at the second step, they find  $\pi_{ci}$  by employing the following identity:

$$\pi_{\delta_i} = \frac{\pi_{ci}}{1 + \pi_{ci}}. \tag{6}$$

Savitsky et al. (2022) use an exact likelihood for the estimation of inclusion probabilities  $\pi_{ci}$ , rather than a pseudo-likelihood based estimation. They propose to stack together nonprobability,  $S_c$ , and probability,  $S_r$ , samples. In this stacked set,  $S$ , indicator variable  $z_i$  takes the value of 1 if unit  $i$  belongs to the nonprobability sample (part 1), and 0 if unit  $i$  belongs to the probability sample (part 2). In this construction, if there is an overlap between the two samples,  $S_c$  and  $S_r$ , then the overlapping units are included into stacked set  $S$  twice: once as a part of the nonprobability sample (with  $z_i = 1$ ) and once as a part of the reference probability sample (with  $z_i = 0$ ). We do not need to know which units overlap or whether there are any overlapping units. The authors use first principles to prove the following relationship between probabilities  $\pi_{z_i}(x_i) = P\{z_i = 1|x_i\}$  of being in part 1 of the stacked set and the sample inclusion probabilities  $\pi_{ci}$  and  $\pi_{ri}$ :

$$\pi_{z_i} = \frac{\pi_{ci}}{\pi_{ri} + \pi_{ci}}. \tag{7}$$

A similar expression (7) was derived by Elliott (2009) and Elliott and Valliant (2017) under the assumption of non-overlapping nonprobability and probability samples. The derivation given in Savitsky et al. (2022) does not require this assumption.

To obtain estimates of  $\pi_{ci}$  from the combined sample, Beresovsky (2019) proposed to parameterize probabilities  $\pi_{ci} = \pi_{ci}(x_i, \theta)$ , as in Chen et al. (2020), and employ identity (7) to present  $\pi_{zi}$  as a composite function of  $\theta$ ; that is,  $\pi_{zi} = \pi_{zi}(\pi_{ci}(x_i, \theta)) = \pi_{ci}(x_i, \theta) / (\pi_{ri} + \pi_{ci}(x_i, \theta))$ .

The log-likelihood for observed Bernoulli variable  $z_i$  is given by

$$\ell_3(\theta) = \sum_{i \in S_c} \log[\pi_{zi}(\pi_{ci}(x_i, \theta))] + \sum_{i \in S_r} \log[1 - \pi_{zi}(\pi_{ci}(x_i, \theta))]. \quad (8)$$

Since the log-likelihood *implicitly* includes a logistic regression model formulation for probabilities  $\pi_{ci}$ , Beresovsky (2019) labeled the proposed approach Implicit Logistic Regression (ILR). For the maximum likelihood estimation (MLE), the score equations are obtained from (8) by taking the derivatives, with respect to  $\theta$ , of the composite function  $\pi_{zi} = \pi_{zi}(\pi_{ci}(\theta))$ . This way, the estimates of  $\pi_{ci}$  are obtained directly from (8) in a single step. Savitsky et al. (2022) parameterized the likelihood, as in (8), and used the Bayesian estimation technique to fit the model.

Note that to implement the ILR approach, the reference sample inclusion probabilities  $\pi_{ri}$  have to be known for all units in the combined set. This is not a limitation for many probability surveys. As discussed in Elliott and Valliant (2017), if probabilities  $\pi_{ri}$  cannot be determined exactly for units in the nonprobability sample, they can be estimated using a regression model. Savitsky et al. (2022) used Bayesian computations to simultaneously estimate  $\pi_{ri}$  and  $\pi_{ci}$  for nonprobability sample units, given available covariates  $x_i$ .

It must be noted that the estimation method of Wang et al. (2021) can be similarly modified to avoid the two-step estimation procedure: a logistic regression model could be formulated for inclusion probabilities  $\pi_{ci}$ , while probabilities  $\pi_{\delta i}$  in (6) could be viewed as a composite function,  $\pi_{\delta i} = \pi_{\delta i}(\pi_{ci}(x_i, \theta)) = \pi_{ci}(x_i, \theta) / (1 + \pi_{ci}(x_i, \theta))$ . This approach is expected to be more efficient. Moreover, it avoids  $\pi_{ci}$  estimates greater than 1 that could occur when the estimation is performed in two steps. Once modified this way, preliminary simulations indicate that Wang et al. (2021) formulation would produce more efficient estimates than the Chen et al. (2020) counterpart, unless in a rare situation where the whole finite population rather than only a reference sample is available.

Simulations show that the exact likelihood method based on formulation of Savitsky et al. (2022) and Beresovsky (2019) performs better than the pseudo-likelihood based alternatives. In the usual situation where the reference probability sample fraction is small, the relative benefits of the exact likelihood approach are even more pronounced.

The existence of a well-designed probability reference sample plays a crucial role in the efforts to reduce the selection bias of a nonprobability sample. Importantly, an ongoing research indicates that the quality of estimates of the nonprobability sample inclusion probabilities is better if there is a good overlap in domains constructed using covariates from both samples. This observation harks back to problems appearing in traditional poststratification methods and to the notion of "representative sampling." Since survey practitioners usually do not have control over the planning or collection of the emerging multitude of nonrandom opportunistic samples, efforts should be directed to developing and maintaining comprehensive probability samples that include sets of good quality covariates. Beaumont et al. (2023)

proposed several model selection methods in application of the modeling nonprobability sample inclusion probabilities.

We now turn our attention to the second data integration situation involving small area estimation, a topic Professor Kalton touched on. This is a problem of great interest for making public policies, fund allocation and regional planning. Small area estimation programs already exist in some national statistical organizations such as the Small Area Income and Poverty Estimates (SAIPE) program of the US Census Bureau (Bell et al., 2016) and Chilean government system (Casas-Cordero Valencia et al., 2016.) The importance placed in the United Nations Sustainable Development Goals (SDG) for disaggregated level statistics is expected to increase the demand for such programs in various national statistical offices worldwide. Standard small area estimation methods generally use statistical models (e.g., mixed models) that combine probability sample data with administrative or census data containing auxiliary variables correlated with the outcome variable of interest. For a review of different small area models and methods, see Jiang and Lahiri (2006), Rao and Molina (2015), Ghosh (2020), and others.

A key to success in small area estimation is to find relevant auxiliary variables not only in the probability sample survey but also in the supplementary big databases. Use of a big probability or nonprobability sample survey could be useful here as surveys typically contain a large number of auxiliary variables that are also available in the probability sample survey. In the context of small area estimation, Sen and Lahiri (2023) considered a statistical data integration technique in which a small probability survey containing the outcome variable of interest is statistically linked with a much bigger probability sample, which does not contain the outcome variable but contains many auxiliary variables also present in the smaller sample. They essentially fitted a mixed model to the smaller probability sample that connects the outcome variable to a set of auxiliary variables and then imputed the outcome variable for all units of the bigger probability sample using the fitted model and auxiliary variables. Finally, they suggested to produce small area estimates using survey weights and imputed values of the outcome variable contained in the bigger probability sample survey. As discussed in their paper, such a method can be used even if the bigger sample is a nonprobability survey using weights constructed by methods such as the ones described earlier.

The development of new approaches demonstrates how the methods of survey estimation continue to evolve by taking into the future the best from fruitful theoretical and methodological developments of the past. As Professor Kalton highlights, we will increasingly encounter data sources that are not produced by standard probability sample designs. Statisticians will find ways to respond to new challenges, as is reflected in the following amusing quote:

...D.J. Finney once wrote about the statistician whose client comes in and says, "Here is my mountain of trash. Find the gems that lie therein." Finney's advice was to not throw him out of the office but to attempt to find out what he considers "gems". After all, if the trained statistician does not help, he will find someone who will....(source: David Salsburg, ASA Connect Discussion)

Of course, nonprobability samples should not be viewed as a “mountain of trash.” Indeed, they can contain a lot of relevant information for producing necessary estimates. It is just that one needs to explore different innovative ways to use information contained in nonprobability samples. In the United States federal statistical system, the need to innovate for combining information from multiple sources has been emphasized in the National Academies of Sciences and Medicine (2017) report on Innovations in Federal Statistics. As discussed, statisticians have been already engaged in suggesting new ideas, such as statistical data integration, to extract information out of multiple non-traditional databases. In coming years, statisticians will be increasingly occupied with finding solutions for obtaining useful information from non-traditional data sources. This is indeed an exciting time for survey statisticians.

## References

- Beaumont, J.-F., K. Bosa, A. Brennan, J. Charlebois, and K. Chu (2023). Handling non-probability samples through inverse probability weighting with an application to statistics canada’s crowdsourcing data. *Survey Methodology* (accepted in 2023 and expected to appear in 2024).
- Bell, W. R., W. W. Basel, and J. J. Maples (2016). *An overview of the US Census Bureau’s small area income and poverty estimates program*, pp. 349–378. Wiley Online Library.
- Beresovsky, V. (2019). On application of a response propensity model to estimation from web samples. In ResearchGate.
- Casas-Cordero Valencia, C., J. Encina, and P. Lahiri (2016). *Poverty mapping for the Chilean Comunas*, pp. 379–404. Wiley Online Library.
- Chen, Y., P. Li, and C. Wu (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association* 115(532), 2011–2021.
- Elliott, M. R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice* 2, 813–845.
- Elliott, M. R. and R. Valliant (2017). Inference for Nonprobability Samples. *Statistical Science* 32(2), 249 – 264.
- Ghosh, M. (2020). Small area estimation: Its evolution in five decades. *Statistics in Transition New Series, Special Issue on Statistical Data Integration*, 1–67.
- Jiang, J. and P. Lahiri (2006). Mixed model prediction and small area estimation, editor’s invited discussion paper. *Test* 15, 1–96.
- Kim J. and K. Morikawa (2023). An empirical likelihood approach to reduce selection bias in voluntary samples. *Calcutta Statistical Association Bulletin* 35 (to appear).

National Academies of Sciences, E. and Medicine (2017). *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington, DC: The National Academies Press.

Rao, J. N. K. and I. Molina (2015). *Small Area Estimation, 2nd Edition*. Wiley.

Savitsky, T. D., M. R. Williams, J. Gershunskaya, V. Beresovsky, and N. G. Johnson (2022). Methods for combining probability and nonprobability samples under unknown overlaps. <https://doi.org/10.48550/arXiv.2208.14541>.

Sen, A. and P. Lahiri (2023). Estimation of finite population proportions for small areas: a statistical data integration approach. <https://doi.org/10.48550/arXiv.2305.12336>.

Wang, L., R. Valliant, and Y. Li (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Stat Med.* 40(4), 5237–5250.



## Discussion of “Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day” by Graham Kalton

Ralf Münnich<sup>1</sup>

Let me first thank Dr. Kalton for his amazing historical review of the development of survey sampling from its origin, contrasting purposive sampling, until now, where some elements of purposive sampling in terms of web or big data seem to supersede the well-elaborated theory of survey statistics. Shall the message be that we do not need any sampling courses at universities anymore, that official statistics should turn to modelling using data with unknown data generating processes, or actually even be substituted by (commercial) *data krakens*? Hardly so! Graham Kalton emphasises a modern thinking about the use of these new data sources which may also have some advantages and he urges future research on data integration methods using (very) different kinds of data while strongly taking quality aspects into account.

Within the last decade, we could observe many new uses of classical data like administrative data and new types of data stemming from internet sources or technical measurement processes such as satellite, mobile phone or scanner data. Already the availability of these new data leads to a huge increase in developing new methodologies and uses. Indeed, official statistics also forced research on new data types, such as scanner data or web-scraped data and others. In Europe, these statistics are often called experimental statistics to emphasise that these statistics cannot (yet) be evaluated using the classical quality concepts, as, e.g. proposed within the European Statistics Code of Practice (<https://ec.europa.eu/eurostat/web/quality/european-quality-standards/european-statistics-code-of-practice>). Some examples can be drawn from [https://www.destatis.de/EN/Service/EXDAT/\\_node.html](https://www.destatis.de/EN/Service/EXDAT/_node.html) or <https://ec.europa.eu/eurostat/web/experimental-statistics>.

During the Covid crisis, and especially in light of the political discussion in Germany, however, one could observe little understanding of data quality and statistics. Timeliness – with its urge of getting data and producing statistics immediately – often lead to the use of available (infection) data, which certainly were influenced by unknown biases. The impact of statistics on these available data in terms of evidence-based policy could hardly be understood at the time, but still legal processes like

---

<sup>1</sup> Economics, Economic and Social Statistics Department, Trier University, Germany.

E-mail: [muennich@uni-trier.de](mailto:muennich@uni-trier.de). ORCID: <https://orcid.org/0000-0001-8285-5667>.



lockdowns were initiated. To state this message more strongly: whenever a legislation process is involved, and especially so if a direct impact on society is the outcome, we must make sure that high quality requirements on data gathering and statistical methodology are set as well as met. High quality typically cannot be achieved with low costs. England was one of the few very good examples during the pandemic, since they were setting up a special Covid survey to better understand the pandemic and to provide adequate and reliable information.

Certainly, this example already shows some critical aspects in data gathering and data quality. Dr. Kalton was emphasising timeliness and accuracy as very important goals of data quality. For sure, these are of utmost importance! However, in practice, both quality principles suffer from budget constraints and cost controls. This directly leads to two questions: Do modern data help to provide more timely and accurate statistics at lower costs? Is there, in case of conflicts, an *ultimate* quality principle?

The first question is already answered by Dr. Kalton. Of course, modern web or big data can help to gather information quickly. Interesting approaches are of course the use of satellite or scanner data. With electronic cash systems, price changes could be tracked much faster than via the use of survey data. However, one always has to understand the advantages as well as the disadvantages of these data generation processes, and one must be able to measure the quality of the output.

Let me briefly sketch one current German debate which, in my view, perfectly fits into this discussion. In the past years, more and more internet surveys were preferred to data from traditional market and opinion research. This immediately led to a discussion on the quality of the outcomes. And certainly, timeliness, accuracy, and costs played an important role within this discussion. The two major arguments were the following: internet surveys suffer from unknown biases. Classical surveys, in the meantime, have to consider response rates considerably below 20%. Under these conditions, most likely both areas have to consider statistical models with strong assumptions to at least reduce possible biases induced by either web surveys or non-response. In my view, one important question has not been raised yet. What is the aim of the survey?

The ultimate aim that necessitates data collection in the first place is of crucial importance for evaluating the importance of the different quality principles. In case one is interested in getting information on current public opinion, probably timeliness and costs are more important than high accuracy. However, in evidence-based policy making, and especially when information for legislative action is needed, I must stress that accuracy must always be considered to be the major principle. This is even more important when large budgets or financial equalization schemes are involved. Additionally, in these cases one must also be able to measure the quality of the outcome of the statistics. This is still a major drawback of using web or big data. And to stress this point, in legislation processes, I strongly urge to involve independent official statistics with its transparent data production process.



With this discussion, I do not want to be misunderstood. Modern data and modern statistical methods are important. And the direction of research, as Dr. Kalton pointed out, will be complex modelling and data integration. Also administrative, register, and related data are important and can provide very good information. However, with all these data, we always have to understand their quality and we should be able to measure the quality of the resulting statistics. Especially in the context of big data, quality measurement may have to be enhanced (cf. Münnich and Articus, 2022, and the citations therein).

Sampling itself may also follow new directions. Classical sampling optimization may be adequately applied in more special cases that allow focusing on specific goals, e.g. the design optimization in the German Censuses 2011 and 2022 (see Münnich et al., 2012, and Burgard, Münnich, and Rupp, 2020). However, likely robustness of methods against assumptions has to be incorporated in design optimization. On the other hand, data integration, multi-source environments, geo-spatial modelling, small area estimation and other modern methods may yield new ideas and directions in sampling theory and application. One example may be sampling from big data sources to reduce complexity.

Despite the mentioned new directions, many ideas have been well-known for a long time. In data analytics, we differentiate between descriptive, predictive, and prescriptive aims. Data that were gathered to describe a state of a system cannot be used to analyse interventions on the system. Indeed, we need the right data and not just merely available data. In conclusion, the exact purpose of the statistics under consideration plays an extremely important role for the selection of data and the priority of the different quality principles.

## References

- Burgard, J. P., Münnich, R., & Rupp, M., (2020). Qualitätszielfunktionen für stark variierende Gemeindegrößen im Zensus 2021. *AStA Wirtschafts- und Sozialstatistisches Archiv*, 14(1), pp. 5–65. With discussion.
- Münnich, R., Articus, C., (2022): Big Data und Qualität – ist viel gleich gut? Pp 85–101. In: Wawrzyniak, B., Herter, M. (Ed.): *Neue Dimensionen in Data Science*. Wichmann.
- Münnich, R., Gabler, S., Ganninger, M., Burgard, J. P.; Kolb, J.-P., (2012): *Stichprobenoptimierung und Schätzung im Zensus 2011*. Destatis: Wiesbaden, Statistik und Wissenschaft, Vol. 21, [https://www.zensus2011.de/SharedDocs/Downloads/DE/Publikationen/Monografien\\_Archiv/2012\\_07\\_Destatis\\_Stochprobenoptimierung\\_und\\_Schaetzung\\_im\\_Zensus.pdf?\\_\\_blob=publicationFile&v=12](https://www.zensus2011.de/SharedDocs/Downloads/DE/Publikationen/Monografien_Archiv/2012_07_Destatis_Stochprobenoptimierung_und_Schaetzung_im_Zensus.pdf?__blob=publicationFile&v=12)



## Rejoinder

Graham Kalton<sup>1</sup>

I should like to thank the discussants for their kind remarks, for their valuable comments on the present state and future directions of the field, and for the many references they cite. Since I have no disagreements with them, I will confine my rejoinder to a few issues that their contributions have surfaced for me.

I will start by rectifying an oversight in my treatment of the early history of survey research and survey sampling: Carl-Erik Särndal has reminded me of the major developments that occurred in Russia during the early years. The impetus for these developments was the need for local self-government units known as *zemstva* to collect data about their populations for administrative purposes. Initially such data were collected with 100% enumerations, but around 1875 sample surveys were introduced for cost savings. The survey procedures were coordinated across *zemstva* and a number of sampling methods were evaluated with input from theoretical statisticians. These statisticians made a number of important contributions, including an impressive early text (1924) entitled *The Foundations of the Theory of the Sampling Method* by A. G. Kowalsky. Although Russian statisticians were at the frontiers of developments in survey sampling until the late 1920's, their contributions were not fully recognized outside Russia. For example, Tschuprow (1923) and Kowalsky in his 1924 text both derived the optimum allocation formula for stratified sampling a decade before Neyman did so in his famous 1934 paper (after learning of Tschuprow's paper, Neyman (1952) recognized Tschuprow's priority for the results). Mespoulet (2002), Zarkovic (1956), Zarkovic (1962), and Seneta (1985) provide further details about early survey research and research on survey sampling in Russia.

Danny Pfeffermann has pointed out that probability samples are almost never representative because of nonresponse—and I would add noncoverage—that is not missing completely at random (NMAR or MCAR). Moreover, I do not think the nonresponse should be viewed as missing at random (MAR), that is MCAR after conditioning on known covariates. Using standard weighting adjustments based on known covariates will not make the sample representative. My favorite quotation from George Box is “Essentially, all models are wrong, but some are useful.” Nonresponse adjustments should be viewed from this perspective as useful but not perfect. Another George Box quotation: “Statisticians, like artists, have the bad habit of falling in love with

---

<sup>1</sup> Joint Program in Survey Methodology, University of Maryland, College Park, MD, USA.

E-mail: [gkalton@gmail.com](mailto:gkalton@gmail.com). ORCID: <https://orcid.org/0000-0002-9685-2616>.



their models.” But there is a difference: artists have artistic license to paint over a model’s blemishes whereas statisticians should attempt to identify and repair the blemishes.

Risto Lehtonen points out the considerable attractions of population registers, as have existed for some time in several Scandinavian countries and are in development elsewhere. Such registers can be viewed as surveys with 100% samples, and the quality of their data should be assessed accordingly: What is their actual coverage? How up-to date are they? How accurate are the data they contain?

Risto’s discussion of population registers also reminded me of a point that I should have addressed more fully: there is a wide variation in the data infrastructure for social research across countries. For example, most developing countries are not in a position to use administrative records or the internet. They rely on probability sample surveys to satisfy their data needs. Fortunately, they have not yet experienced the severe declines in response rates that are so harmful to surveys in most high-income countries.

Julie Gershunskaya and Partha Lahiri address two important current areas of research. One is the research on how to employ a probability sample to reduce the bias in estimates from a nonprobability sample, making use of auxiliary variables collected in both samples. The auxiliary variables aim to capture the key variables that are predictors of membership in the nonprobability sample. Challenges to be addressed with this approach include identifying the key variables; dealing with the fact that some response categories that occur frequently in the probability sample are very sparsely represented in the nonprobability sample; and concerns about the equivalence of the responses to the key variables obtained in the two samples that use different modes of collection. The results from this approach should be viewed with caution. However, recalling George Box’s quotation above, imperfect models can be useful. Julie and Partha rightly say that the aim of these models is to reduce, not eliminate, bias. The question to be asked is how to assess whether the models have reduced bias to an acceptable level.

The second area that Julie and Partha address is small area estimation. I should have written more about this methodology whose use has now become so widespread. My first practical exposure to small area estimation occurred in the late 1990’s, when I chaired a National Academy of Sciences’ panel that was asked to advise about the quality of the small area estimates of the numbers of poor school-aged children that were being developed in the U.S. Census Bureau’s Small Area Income and Poverty Estimates (SAIPE) program. The central issue was whether the estimates, which were produced for 3,000 counties and 14,000 school districts, were appropriate and sufficiently reliable to be used in allocating very large sums of money directly to school districts. At that time, this was a novel application of small area estimates, and subject to considerable questioning. After extensive evaluation of the area level models by both the Panel and the Census Bureau (Citro and Kalton, 2000), the Panel concluded that the small area estimates were “fit for use” for the purpose of this fund allocation, despite a recognition of substantial errors in the individual estimates. The Panel was influenced by the fact that the legislation stipulated that the funds should be distributed directly to the school districts and that, even though the small area estimates were not ideal, they were the best available. I was persuaded by my experience on the Panel that, with strong predictors and careful model

development and testing, small area estimation methods have an important role to play in responding to policy makers' increasing demands for local area estimates.

Ralf Münnich emphasizes the importance of assessing the overall quality of statistical estimates in the light of the uses of the estimates. As he notes, timeliness is often in conflict with accuracy. In some situations, timeliness may be paramount, and accuracy may suffer. However, one must guard against the risk that accuracy is so low that the resulting estimates are misleading. Estimates based on big data sources or even large surveys conducted with an overriding emphasis on speed may, because of their sample sizes, appear to be well-grounded but that may well be illusory.

It is often argued that although individual estimates may be subject to serious biases, these biases will cancel out for differences between estimates, either between subgroups of the sample or across time. While the underlying model for that argument often appears reasonable, the assumptions underpinning it need to be carefully assessed in each case.

Ralf also points out the importance of cost constraints. When the cost constraints severely limit a study to a very small sample size, it may be preferable to forego the extra costs involved in selecting and fielding a probability sample, in favor of a quasi-probability sample or a nonprobability sample design. As Kish (1965, p. 29) notes: "Probability sampling is not a dogma, but a strategy, especially for large numbers."

Finally, Ralf and other discussants have pointed out the attractions of data integration. I also see these attractions, but I think that the challenges of mode effects arising from different data sources should not be underestimated.

***In conclusion, I congratulate Statistics in Transition on celebrating its 30<sup>th</sup> anniversary. It plays a distinct and important role among statistics journals. With the major changes in statistical methodology taking place in official statistics and in social research, it has a bright future for the contributions it can make.***

## References

- Citro, C. F., Kalton, G. Eds., (2000). *Small-Area Estimates of School-Age Children in Poverty: Evaluation of Current Methodology*. National Academy Press, Washington D.C.
- Kish, L., (1965). *Survey Sampling*. Wiley, New York.
- Mespoulet, M., (2002). From typical areas to random sampling: sampling methods in Russia from 1875 to 1930. *Science in Context*, 15(3), pp. 411–425.
- Neyman, J., (1952). Recognition of priority. *Journal of the Royal Statistical Society, A*, 115(4), 602.
- Seneta, E., (1985). A sketch of the history of survey sampling in Russia. *Journal of the Royal Statistical Society, A*, 148(2), pp. 118–125.
- Tschuprow, A. A., (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron*, 2(4), pp. 646–683.
- Zarkovic, S. S., (1956). Note on the history of sampling methods in Russia. *Journal of the Royal Statistical Society, A*, 119(3), pp. 336–338.
- Zarkovic, S. S., (1962). A supplement to "Note on the history of sampling methods in Russia". *Journal of the Royal Statistical Society, A*, 125(4), pp. 580–582.



# Rotation schemes and Chebyshev polynomials

Jacek Wesolowski<sup>1</sup>

## Abstract

There is a continuing interplay between mathematics and survey methodology involving different branches of mathematics, not only probability. This interplay is quite obvious as regards the first of the two options: probability vs. non-probability sampling, as proposed and discussed in Kalton (2023). There, mathematics is represented by probability and mathematical statistics. However, sometimes connections between mathematics and survey methodology are less obvious, yet still crucial and intriguing. In this paper we refer to such an unexpected relation, namely between rotation sampling and Chebyshev polynomials. This connection, introduced in Kowalski and Wesolowski (2015), proved fundamental for the derivation of an explicit form of the recursion for the BLUE  $\hat{\mu}_t$  of the mean on each occasion  $t$  in repeated surveys based on a cascade rotation scheme. This general result was obtained under two basic assumptions: ASSUMPTION I and ASSUMPTION II, expressed in terms of the Chebyshev polynomials. Moreover, in that paper, it was conjectured that these two assumptions are always satisfied, so the derived form of recursion is universally valid. In this paper, we partially confirm this conjecture by showing that ASSUMPTION I is satisfied for rotation patterns with a single gap of an arbitrary size.

## 1. Introduction

Existence of connections between survey methodology and mathematics is a trivial statement. The most natural ones are triggered by probability sampling, the first option in dichotomy between probability and non-probability sampling proposed in the review, Kalton (2023), in this issue of SiT. Of course, it involves probability theory and mathematical statistics on the mathematical side. The title (and the content) of the popular monograph "Model Assisted Sampling Survey" by Särndal, Swensson and Wretman (1992) is the best reference to appreciate this connection. Some other areas of mathematics are also typically involved in this interplay; as (convex) optimization theory in optimal allocation problems, or graph theory in modelling dependence structure in adaptive sampling. The second part of Kalton's dichotomy may open new doors for involvement of mathematics in survey methodology. But even within survey methodology based on probability sampling, unexpected and useful connections between the two areas happen. A good example is a connection between rotation sampling and Chebyshev polynomials, which we are going to explore in this paper.

Rotation of the sample is a standard method used in repeated surveys. It allows not only catching the dynamics of the population under study and lower the burden of surveys for respondents, but also can be used to improve estimation of parameters at the given occasion

<sup>1</sup>Statistics Poland and Warsaw University of Technology, Poland. E-mail: [jacek.wesolowski@pw.edu.pl](mailto:jacek.wesolowski@pw.edu.pl).

ORCID: <https://orcid.org/0000-0001-7615-694X>.

© Jacek Wesolowski. Article available under the CC BY-SA 4.0 licence



by proper treatment of observations from the past occasions. Typical examples are the Labour Force Survey in the EU with the rotation pattern 110011 (also referred to by 2-2-2), i.e. a unit (group of units) is in a sample for two consecutive occasions, leaves the survey for next two occasions, then enters the sample for two more consecutive occasions and then leaves the survey for good, or the Current Population Survey in the US with the rotation pattern 111100000001111 (i.e. 4-8-4). Such methodology was proposed in the seminal paper Patterson (1950), who postulated the recurrence form for the best linear unbiased estimators (BLUEs) of the mean on each occasion. Patterson considered a model with exponentially time-dependent correlations for each unit of the population and independence between units. He assumed that the rotation pattern is such that any unit leaving the sample cannot return to the survey. In such setting it was proved that for any occasion  $t$  the BLUE  $\hat{\mu}_t$  (based on all past observations) of the current mean  $\mu_t$  satisfies the linear one-step recursion of the form

$$\hat{\mu}_t = a_1(t)\hat{\mu}_{t-1} + r_0^T(t)\underline{X}_t + r_1^T(t)\underline{X}_{t-1}, \quad (1)$$

where  $\underline{X}_i$  is the vector of observations at time  $i = t, t-1$  and the recursion coefficients, i.e. the number  $a_1(t)$  and the vectors  $r_0(t)$ ,  $r_1(t)$  were identified in terms of the correlation coefficient  $\rho$ .

The assumption that a unit leaving the sample never returns to the survey was crucial for derivation of (1). Therefore, it was expected that the first order recursion for the optimal BLUE's would no longer hold for more general rotation patterns which do not satisfy Patterson's condition. A postulated form of the recursion would be of the form

$$\hat{\mu}_t = a_1(t)\hat{\mu}_{t-1} + \dots + a_p(t)\hat{\mu}_{t-p} + r_0^T(t)\underline{X}_t + r_1^T(t)\underline{X}_{t-1} + \dots + r_p^T(t)\underline{X}_{t-p}, \quad (2)$$

where  $p$  is a natural number and  $a_1(t), \dots, a_p(t)$ ,  $r_0(t), \dots, r_p(t)$  are numeric and vector coefficients. However, such extension posed major difficulties, see, e.g. Yansaneh and Fuller (1998). Therefore, for years researchers have been mostly focused on sub-optimal estimators. Already Hansen, Hurwitz, Nisselson and Steinberg (1955) proposed an alternative sub-optimal  $K$ -composite estimator, where the optimality was sought under additional assumption of one-step recursion, that is, under assumption that  $p = 1$  in (2). This approach was further developed in Rao and Graham (1964), Gurney and Daly (1965), Cantwell (1990), Cantwell and Caldwell (1998), Ciepela, Gniado, Wesołowski and Wojtyś (2012). Another approach, based on the so-called regression composite estimator has been proposed and studied in Bell (2001), Fuller and Rao (2001), Singh, Kennedy and Wu (2001), Kowalczyk and Juszcak (2018). Different rotation patterns and comparisons of efficiencies of different methods are presented in McLaren and Steel (2000), and Steel and McLaren (2002,2008). For a relatively new review see Karna and Nath (2015). Polish experiences with rotation sampling are described in a review by Kordos (2012). An alternative methodology, which we do not consider here, is based on time series theory, with random means on subsequent occasions while here we assume that they are constants depending on  $t$ . An overview of the time series approach to rotation sampling is given e.g. in Binder and Hidirogrou (1988).

The first result going beyond Patterson's scheme of a rotation pattern without gaps, i.e. of the form 11...11, was obtained in Kowalski (2009), where it was proved that for



rotation patterns with arbitrary number of singleton gaps, i.e. of the form  $1...101...101...1$ , the recursion (2) holds with  $p = 2$  and all coefficients were identified. Moreover, it was observed in that paper that the coefficients stabilize quickly as  $t$  grows, which suggested an approach to the general case by recursion with coefficients not depending on  $t$ , equivalently for the stationary situation, i.e. the case when  $t \rightarrow \infty$ . Then, the recursion assumes the form

$$\hat{\mu}_t = a_1 \hat{\mu}_{t-1} + \dots + a_p \hat{\mu}_{t-p} + \underline{r}_0^T \underline{X}_t + \underline{r}_1^T \underline{X}_{t-1} + \dots + \underline{r}_p^T \underline{X}_{t-p}, \tag{3}$$

Under such a setting a general solution for arbitrary rotation pattern was obtained in Kowalski and Wesołowski (2015) (referred to by KW in the sequel). According to the main result in KW the recursion depth,  $p$ , is the size of the maximal gap in the rotation pattern increased by 1 (therefore it was 1 in the Patterson model, 2 in for rotation patterns with gaps of size 1) and 3 in the LFS rotation pattern 110011 (the last one settled in Wesołowski (2010)). The form of the coefficients in (3), as given in KW, is explicit, and rather unexpectedly, involves the Chebyshev polynomials of the first kind defined by

$$T_k(x) = \cos(k \arccos x), \quad k = 0, 1, \dots$$

For a thorough review of Chebyshev polynomials readers are encouraged to consult Paszkowski (1975). It has to be emphasized that the solution, valid for any cascade rotation pattern, was obtained in KW under two specific assumptions: ASSUMPTION I concerning roots of a special polynomial  $Q_p$  of degree  $p$  expressed through Chebyshev polynomials and ASSUMPTION II concerning full rank of certain matrix  $S$  being a function of these roots. However, in numerous simulations both these ASSUMPTIONS were always satisfied. Therefore, it was conjectured, see p. 101 of KW, that both ASSUMPTIONS are always satisfied and the solution obtained is universally valid. The goal of the present paper is to show that the conjecture holds true, at least partially. Actually, it will be shown that ASSUMPTION I holds true for rotation patterns with a single gap of arbitrary size. The rest of the paper is organized as follows. In Section 2, we present the general setting of the rotation scheme in mathematical language and adjust ASSUMPTIONS I and II to rotation patterns with a single gap of arbitrary size. In Section 3, we give a short introduction to Chebyshev polynomials emphasizing tools needed to analyze roots of the polynomial  $Q_p$ . In Section 4, we prove the main result which says that ASSUMPTION I is satisfied for rotation patterns with single gap of arbitrary size. Section 5 is devoted to a representation of  $Q_p$  as an affine perturbation of a Chebyshev polynomial of a proper degree, which is the main tool for the proof.

## 2. General setting and rotation patterns with a single gap of arbitrary size

Consider a doubly-infinite matrix of random variables  $(X_{ij})$ ,  $i, j \in \mathbb{Z}$  such that for any  $j \in \mathbb{Z}$

$$\mathbb{E}X_{i,j} = \mu_j, \quad \text{for all } i \in \mathbb{Z},$$

and, without loss of generality we assume that  $\text{Var}(X_{i,j}) = 1$  for all  $i, j \in \mathbb{Z}$ . The correlation structure is described by

$$\text{Corr}(X_{i,j}, X_{k,l}) = I(k=i)\rho^{|j-l|}, \quad (4)$$

where  $0 < |\rho| < 1$ .

For natural number  $N \geq 2$  consider a sequence  $\underline{X}_j = (X_{j,j}, \dots, X_{j+N-1,j})$ ,  $j \in \mathbb{Z}$ , of  $N$ -variate random vectors. Note that from (4) it follows that the covariance matrix  $\mathbf{C} = \text{Cov}(\underline{X}_j, \underline{X}_{j+1})$ , of dimensions  $N \times N$ , has all entries equal zero except the ones just above the diagonal, which are all equal  $\rho$ . Moreover, (4) yields

$$\text{Cov}(\underline{X}_j, \underline{X}_k) = \mathbf{C}^{|k-j|}$$

and note that  $\mathbf{C}^j$  is a matrix with all entries equal zero except the  $j$ th over diagonal with all entries equal  $\rho^j$  when  $j \leq N-1$  and it is a zero matrix when  $j > N-1$ .

A rotation pattern is any vector  $(\varepsilon_1, \dots, \varepsilon_N)$  with 0-1 entries such that  $\varepsilon_1 = \varepsilon_N = 1$ . Let  $M = \{j \in \{1, \dots, N\} : \varepsilon_j = 0\}$ . Then  $N = n + m$ , where  $m = \#M$  is the number of zeros among the entries and  $n$  is the number of ones (note that  $n \geq 2$ ). Each zero in rotation pattern results in a "hole" in the sample and the largest set of subsequent zeros determines a gap in the rotation pattern. Let  $p-1$  denote the dimension of the largest gap in the rotation pattern.

We modify vectors  $\underline{X}_j$  into

$$\underline{Y}_j = (X_{j+k-1,j}, k \in \{1, \dots, N\} \setminus M), \quad j \in \mathbb{Z}.$$

For a given  $t \in \mathbb{Z}$  let  $\hat{\mu}_t$  denote the BLUE of  $\mu_t$  based on  $\underline{Y}_s$ ,  $s \leq t$ .

We study the recurrence formula for the BLUE estimators of the following form

$$\hat{\mu}_t = \tilde{a}_1 \hat{\mu}_{t-1} + \dots + \tilde{a}_s \hat{\mu}_{t-s} + \tilde{r}_0^T \underline{Y}_t + \tilde{r}_1^T \underline{Y}_{t-1} + \dots + \tilde{r}_s^T \underline{Y}_{t-s},$$

for any  $t \in \mathbb{Z}$ , where  $s, \tilde{a}_1, \dots, \tilde{a}_s \in \mathbb{R}$  and  $\tilde{r}_0, \tilde{r}_1, \dots, \tilde{r}_s \in \mathbb{R}^m$  are unknown. The goal is to find  $s$  and to identify remaining parameters in terms of  $p, \rho$  and  $N$ .

Alternatively,  $\hat{\mu}_t$  can be defined as optimal unbiased linear estimator  $\sum_{s \leq t} w_s^T \underline{X}_s$ , with additional constraints

$$w_{s,j}(1 - \varepsilon_j) = 0, \quad j = 1, \dots, N, \quad s \leq t, \quad (5)$$

imposed by the gaps in the rotation pattern. Therefore, the above recursion can be written in the form

$$\hat{\mu}_t = a_1 \hat{\mu}_{t-1} + \dots + a_s \hat{\mu}_{t-s} + \underline{r}_0^T \underline{X}_t + \underline{r}_1^T \underline{X}_{t-1} + \dots + \underline{r}_s^T \underline{X}_{t-s}, \quad (6)$$

for any  $t \in \mathbb{Z}$ , where  $a_1, \dots, a_s \in \mathbb{R}$  and  $\underline{r}_0, \underline{r}_1, \dots, \underline{r}_s \in \mathbb{R}^N$ .

Note that (5) forces respective entries of vectors  $\underline{r}_j \in \mathbb{R}^N$ ,  $j = 0, \dots, s$ , to be equal zero.

The problem is to prove that the recurrence (6) holds for  $s = p$  and to determine scalar parameters  $a_i, i = 1, \dots, p$  and vector parameters  $r_j \in \mathbb{R}^N, j = 0, 1, \dots, p$ . As it has been already mentioned, under two basic assumptions there exist formulas which completely answer this question. The first of these assumptions is concerned with localization of roots of certain polynomial and the second deals with unique solvability of certain linear system of equations. There is a strong numerical evidence that these assumptions may be universally satisfied. However, no proof of this fact has been available until now. It has been theoretically confirmed only for  $m = 0, 1$  and any  $n \geq 2$  and for the rotation pattern 110011 (here  $m = 2$ ). In this paper we will show that the first assumption (ASSUMPTION I below) is satisfied for all rotation patterns with a single gap of arbitrary size  $m$ . We do not know how to prove that the second assumption (ASSUMPTION II below) is satisfied in this case.

From now on we consider only rotation patterns with a single gap of arbitrary size  $m \in \{0, 1, \dots\}$ . In the remaining part of this section we will present ASSUMPTIONS I and II for such rotation patterns only. A reader interested in the general case is encouraged to look into KW.

Recall that the Chebyshev polynomials of the first kind ( $T_n$ ) are defined through a three step recurrence

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad n = 1, 2, \dots \tag{7}$$

and  $T_0(x) = 1, T_1(x) = x$ , that is  $T_n(\cos t) = \cos(nt), n = 0, 1, \dots$

Consider a polynomial  $Q_p$  of degree  $p$  defined by

$$Q_p(x) = 1 - \rho^2 + (N - 1)(1 + \rho^2 - 2\rho x) - (1 + \rho^2 - 2\rho x)^2 \operatorname{tr}(\mathbf{T}_m(x)\mathbf{R}_m^{-1}(\rho)), \tag{8}$$

where  $\mathbf{T}_m$  is an  $m \times m$  symmetric Toeplitz matrix of the Chebyshev polynomials of the form

$$\mathbf{T}_m = \begin{bmatrix} T_0 & T_1 & T_2 & \dots & T_{m-2} & T_{m-1} \\ T_1 & T_0 & T_1 & \dots & T_{m-3} & T_{m-2} \\ \vdots & \vdots & \vdots & \dots & \vdots & \\ T_{m-2} & T_{m-3} & T_{m-4} & \dots & T_0 & T_1 \\ T_{m-1} & T_{m-2} & T_{m-3} & \dots & T_1 & T_0 \end{bmatrix} \tag{9}$$

and  $\mathbf{R}_m$  is an  $m \times m$  invertible constant three-diagonal matrix

$$\mathbf{R}_m = \begin{bmatrix} 1 + \rho^2 & -\rho & 0 & \dots & 0 & 0 \\ -\rho & 1 + \rho^2 & -\rho & \dots & 0 & 0 \\ 0 & -\rho & 1 + \rho^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \\ 0 & 0 & 0 & \dots & 1 + \rho^2 & -\rho \\ 0 & 0 & 0 & \dots & -\rho & 1 + \rho^2 \end{bmatrix} \tag{10}$$

**ASSUMPTION I: Roots of  $Q_p$  are distinct and do not belong to  $[-1, 1]$ .**

As mentioned above, our goal is to show that ASSUMPTION I is satisfied. It is done in the remaining three sections below. In Section 3, we present some basic facts on the Chebyshev polynomials of the first and second kind we need in the sequel. The proof, given in Section 4, to large extent is based on a representation of  $Q_p$  derived in Section 5.

But before we analyze ASSUMPTION I we will introduce also ASSUMPTION II, which is conjectured to be also satisfied, but we do not know, how to prove it.

Note that  $Q_p$  is a polynomial of  $p$ th degree. If its roots  $x_1, \dots, x_p$  are simple and are outside of the interval  $[-1, 1]$  (which will be proved in the sequel), then there exist unique  $d_1, \dots, d_p$ , which can be complex, such that  $|d_i| < 1$  and  $\frac{1}{2}(d_i + d_i^{-1}) = x_i, i = 1, \dots, p$ .

For such numbers  $d_1, \dots, d_p$  define a  $p^2 \times p^2$  matrix

$$\mathbf{S} = \mathbf{S}(d_1, \dots, d_p) = \begin{bmatrix} \tilde{\mathbf{G}}(d_1) & \tilde{\mathbf{G}}(d_2) & \cdots & \tilde{\mathbf{G}}(d_p) \\ \mathbf{G}(d_1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{G}(d_2) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G}(d_p) \end{bmatrix}$$

where  $\tilde{\mathbf{G}}(d_i)$  are  $p \times p$  matrices

$$\tilde{\mathbf{G}}(d) = \frac{1}{1-\rho^2} \begin{bmatrix} (N-1)(1-d\rho) + 1 - \rho^2 & (1-d\rho)\mathbf{1}_h^T \\ (1-d\rho)\mathbf{1}_{p-1} & \tilde{\mathbf{H}}_{p-1} \end{bmatrix}$$

with  $\tilde{\mathbf{H}}_{p-1}(d)$  being a  $(p-1) \times (p-1)$  upper bi-diagonal matrix

$$\tilde{\mathbf{H}}_{p-1}(d) = \begin{bmatrix} 1 & -d\rho & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & \ddots & -d\rho \\ & & & & 1 \end{bmatrix}.$$

and  $\mathbf{G}(d_i)$  are  $(p-1) \times p$  matrices

$$\mathbf{G}(d) = \frac{1}{1-\rho^2} [(1-d\rho)(d-\rho)\mathbf{1}_h, d\mathbf{H}_{p-1}],$$

with  $\mathbf{H}_{p-1} = \mathbf{H}_{p-1}(d)$  being a  $(p-1) \times (p-1)$  tri-diagonal matrix

$$\mathbf{H}_{p-1}(d) = \begin{bmatrix} 1+\rho^2 & -d\rho & & & \\ -\rho/d & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & -\rho/d & -d\rho & \\ & & & -\rho/d & 1+\rho^2 \end{bmatrix}.$$

Here is the second main assumption:

ASSUMPTION II:  $\det \mathbf{S}(\mathbf{d}_1, \dots, \mathbf{d}_p) \neq 0$ .

Unfortunately, we are unable to prove that it is satisfied in the setting of a single gap of arbitrary size  $m$ . As mentioned above, only the cases of  $m = 0, 1, 2$  have been settled until now.

When ASSUMPTION I and ASSUMPTION II are satisfied, then Theorem 3.1 proved in KW says that  $p = m + 1$  and gives explicit formulas for  $a_i, i = 1, \dots, p$ , and  $r_i, i = 0, 1, \dots, p$ , in terms of the  $d_1, \dots, d_p$  determined through roots of  $Q_p$  and the solution  $\underline{c}$  of the linear equation  $\mathbf{S}\underline{c} = (1, 0, \dots, 0)^T$ . For details consult KW.

### 3. Chebyshev polynomials

The Chebyshev polynomials of the second kind  $(U_n)_{n \geq 0}$  are defined through the same three step recurrence as  $(T_n)_{n \geq 0}$ , that is

$$U_{n+1}(x) = 2xU_n(x) - U_{n-1}(x), \quad n = 1, 2, \dots \tag{11}$$

but the boundary conditions are slightly different:  $U_0(x) = 1$  and  $U_1(x) = 2x$ , that is  $U_n(\cos t) = \frac{\sin((n+1)t)}{\sin t}$ , if only  $\sin t \neq 0$ .

We will also use two important identities connecting two forms of the Chebyshev polynomials for any  $n = 1, 2, \dots$  (in the formulas below we denote  $U_{-1} \equiv 0$ ):

$$T'_n = nU_{n-1}, \tag{12}$$

and

$$T_n^2(x) + (1 - x^2)U_{n-1}^2(x) = 1. \tag{13}$$

Moreover, two representations of the Chebyshev polynomials given in Lemma 3.1 (cf. Paszkowski, 1975) below will be very useful.

**Lemma 3.1.** For any  $x \neq 0$  and  $n = 0, 1, \dots$

$$T_n\left(\frac{1}{2}(x + x^{-1})\right) = \frac{1}{2}(x^n + x^{-n}) \tag{14}$$

and for  $x \neq 0, \pm 1$  we have

$$U_n\left(\frac{1}{2}(x + x^{-1})\right) = \frac{x^{n+1} - x^{-(n+1)}}{x - x^{-1}}. \tag{15}$$

It is known (cf. Paszkowski, 1975) that

$$U_n(x) = \det \mathbf{V}_n(x), \quad n = 1, 2, \dots, \tag{16}$$

where  $\mathbf{V}_n(x)$  is an  $n \times n$  tridiagonal matrix defined by

$$\mathbf{V}_n(x) = \begin{bmatrix} 2x & -1 & 0 & \dots & 0 & 0 \\ -1 & 2x & -1 & \dots & 0 & 0 \\ 0 & -1 & 2x & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \\ 0 & 0 & 0 & \dots & 2x & -1 \\ 0 & 0 & 0 & \dots & -1 & 2x \end{bmatrix}. \tag{17}$$

We see that  $\mathbf{V}_n(x)$  is non-singular for any  $x \geq 1$ . In this case the explicit form of the inverse of  $\mathbf{V}_n(x)$  is known. Let  $\mathbf{A}$  denote the inverse of  $\mathbf{V}_n(x)$ . Then  $\mathbf{A} = [a_{i,j}]_{i,j \in \{1, \dots, n\}}$  is a symmetric matrix such that

$$a_{i,j} = \frac{1}{U_n(x)} U_{i-1}(x) U_{n-j}(x), \quad 1 \leq i \leq j \leq n. \quad (18)$$

We will apply the following useful formulae (cf. Paszkowski, 1975)

$$\sum_{j=1}^n T_j(x) U_{n-j}(x) = \frac{n}{2} U_n(x), \quad (19)$$

$$2(x-y) \sum_{j=0}^n {}'T_j(x) U_{n-j}(y) = T_{n+1}(x) - T_{n+1}(y), \quad (20)$$

under notation

$$\sum_{j=s}^n {}'b_j = \frac{1}{2} b_s + b_{s+1} + \dots + b_n, \quad n > l.$$

#### 4. Roots of $Q_p$

In this section we show that ASSUMPTION 1 is satisfied for rotation patterns with a single gap of an arbitrary size. In the proof we strongly rely on properties of the Chebyshev polynomials and the representation of  $Q_p$  in terms of an affine additive perturbation of the Chebyshev polynomial of the first kind derived in Section 5.

**Theorem 4.1.** *For any  $p \geq 1$  the polynomial  $Q_p$  defined by (8), (9) and (10) has exactly one (when  $p$  is odd) or exactly two (when  $p$  is even) real roots (i.e. the remaining roots are complex). These roots are outside of interval  $[-1, 1]$ . All roots of  $Q_p$  are simple.*

*Proof.* Note that due to Prop. 5.2

$$(\det \mathbf{R}_m) Q_p(x) = \det \mathbf{R}_m (n-2)(1+\rho^2 - 2\rho x) + 2 - 2\rho^{m+1} T_{m+1}(x). \quad (21)$$

Therefore, the roots of  $Q_p$  are identical to the roots of polynomial  $\tilde{Q}_p$  defined by

$$\tilde{Q}_p(x) = a + bx + T_{m+1}(x),$$

where

$$a = -\frac{(\det \mathbf{R}_m)(n-2)(1+\rho^2)+2}{2\rho^{m+1}} = -rU_m(r)(n-2) - \rho^{-m-1} \quad \text{and} \quad b = \frac{(\det \mathbf{R}_m)(n-2)}{\rho^m} = U_m(r)(n-2).$$

Assume that  $z_0$  is a multiple root of  $\tilde{Q}_p$ . That is

$$a + bz_0 + T_{m+1}(z_0) = 0. \quad (22)$$

Moreover,  $z_0$  is necessarily a root of derivative of  $\tilde{Q}_p$ . Thus from (12) we get

$$b + (m+1)U_m(z_0) = 0. \quad (23)$$

Combining (22) and (23) through (13) we obtain

$$(a + bz_0)^2 + (1 - z_0^2) \left( \frac{b}{m+1} \right)^2 = 1.$$

That is,  $z_0$  is a solution of the quadratic equation

$$b^2 \left( 1 - \frac{1}{(m+1)^2} \right) x^2 + 2abx + a^2 + \left( \frac{b}{m+1} \right)^2 - 1 = 0. \tag{24}$$

whose discriminant is

$$\Delta = 4b^2 \left[ \frac{a^2}{(m+1)^2} - \left( \frac{b^2}{(m+1)^2} - 1 \right) \left( 1 - \frac{1}{(m+1)^2} \right) \right].$$

If  $b \leq m + 1$  clearly  $\Delta > 0$ . For  $b > m + 1$  note that

$$a^2 = \frac{[(\det \mathbf{R}_m)(n-2)(1+\rho^2)+2]^2}{4\rho^{2m+2}} > \frac{[(\det \mathbf{R}_m)(n-2)(1+\rho^2)]^2}{4\rho^{2m+2}} = b^2 \left( \frac{1+\rho^2}{2\rho} \right)^2 > b^2.$$

Therefore,  $\Delta > 0$  also in this case. Thus, the quadratic equation (24) has only real solutions. Consequently,  $\tilde{Q}_p$  does not have multiple complex roots.

Note that  $\tilde{Q}_p$  can be written as

$$\tilde{Q}_p(x) = -\frac{1}{\rho^{m+1}} \left[ 1 + \frac{1}{2}(n-2)(1+\rho^2-2\rho x) \det \mathbf{R}_m \right] + T_{m+1}(x).$$

Clearly, the expression in brackets is greater or equal 1 for  $x \in [-1, 1]$ . Since the Chebyshev polynomials  $T_n, n = 1, 2, \dots$ , on  $[-1, 1]$  assume values in  $[-1, 1]$  it follows that on  $[-1, 1]$  the polynomial  $\tilde{Q}_p$  is either strictly positive (when  $\rho^{m+1} < 0$ ) or strictly negative (when  $\rho^{m+1} > 0$ ).

It is well known that

- if  $m$  is an even number then:  $U_m$  is strictly decreasing on  $(-\infty, -1)$ , strictly increasing on  $(1, \infty)$  and  $U_m(\pm 1) = m + 1$ ;
- if  $m$  is an odd number then:  $U_m$  is strictly increasing on  $(-\infty, -1)$  and on  $(1, \infty)$  and  $U_m(\pm 1) = \pm(m + 1)$ .

Consequently, only the following four cases are possible:

1. If  $m$  is even and  $\rho > 0$  then  $a < -1$  and  $b \geq 0$ . Thus,  $\tilde{Q}_p$  (of odd degree) has exactly one real root  $x_1 > 1$ . Note that it is simple. The reason for that is that the derivative of  $\tilde{Q}_p$  which equals  $b + (m + 1)U_m(x)$  is bounded from below by  $b + (m + 1)^2$  on  $(1, \infty)$ . Therefore,  $\tilde{Q}_p$  cannot have a multiple real root  $> 1$ .
2. If  $m$  is even and  $\rho < 0$  then  $a > 1$  and  $b \geq 0$ . Thus,  $\tilde{Q}_p$  (of odd degree) has exactly one real root  $x_1 < -1$ . Similarly, as above  $\tilde{Q}'_p(x) = b + (m + 1)U_m(x) > b + (m + 1)^2$  on  $(-\infty, -1)$ , and thus  $\tilde{Q}_p$  does not have a multiple root  $< -1$ .
3. If  $m$  is odd and  $\rho > 0$  then  $a < -1$  and  $b \geq 0$ . Thus,  $\tilde{Q}_p$  (of even degree) has exactly two real roots:  $x_1 < -1$  and  $x_2 > 1$ . Similarly as above  $\tilde{Q}'_p(x) = b + (m + 1)U_m(x) > b + (m + 1)^2 > 0$  for  $x > 1$ , and thus the root  $x_2$  is simple. Note also that the quadratic polynomial (24) is strictly positive on negative half line, that is  $\tilde{Q}_p$  cannot have negative multiple roots, in particular, the root  $x_1$  is not multiple.
4. If  $m$  is odd and  $\rho < 0$  then  $a < -1$  and  $b \leq 0$ . Thus,  $\tilde{Q}_p$  (of even degree) has exactly two real roots:  $x_1 < -1$  and  $x_2 > 1$ . This time the derivative,  $\tilde{Q}'_p(x) = b + (m + 1)U_m(x) < b - (m + 1)^2 < 0$  for  $x < -1$ , and thus the root  $x_1$  is simple. Similarly as above, to check that  $x_2$  is simple, we refer to (24) having the left-hand side strictly positive for  $x > 0$ , which means that there are no multiple positive roots.

□

**Remark 4.1.** Note that from (21) for  $n = 2$  we get  $(\det \mathbf{R}_m) Q_p(x) = 2 - 2\rho^{m+1} T_{m+1}(x)$ . Consequently, to find roots of  $Q_p$  it suffices to look for solutions of the equation

$$T_{m+1}(x) = \rho^{-m-1}.$$

For  $x = \frac{1}{2}(d + 1/d)$  we obtain

$$d^{m+1} + d^{-m-1} = \frac{2}{\rho^{m+1}}$$

and thus for  $z = d^{m+1}$  we get a quadratic equation

$$z^2 - 2\frac{z}{\rho^{m+1}} + 1 = 0$$

with two real solutions

$$z = \frac{1 \pm \sqrt{1 - \rho^{2(m+1)}}}{\rho^{m+1}}.$$

Therefore

$$d_j = d_{\pm} \exp \left[ i \frac{2j\pi}{m+1} \right], \quad j = 0, 1, \dots, m,$$

where

$$d_{\pm} = \frac{m+1 \sqrt{1 \pm \sqrt{1 - \rho^{2(m+1)}}}}{|\rho|}.$$

Note that  $0 < d_- < 1 < d_+$ .

## 5. $Q_p$ through additive first degree perturbation of $T_p$

In this section we derive a convenient representation of  $Q_p$  in terms of  $T_p$  with changed terms of degree zero and one. It is preceded by a simple expression for determinant of  $\mathbf{R}_m$  involving the second order Chebyshev polynomial  $U_m$ .

**Lemma 5.1.** Let  $r = \frac{1}{2} \left( \rho + \frac{1}{\rho} \right)$ . For any  $m = 0, 1, \dots$

$$\det \mathbf{R}_m = \rho^m U_m(r), \quad m = 0, 1, 2, \dots \quad (25)$$

*Proof.* Notice that  $\mathbf{R}_m = \rho \mathbf{V}_m(r)$ , so by (16) we have

$$\det \mathbf{R}_m = \rho^m \det \mathbf{V}_m(r) = \rho^m U_m(r).$$

The proof is complete. □

**Proposition 5.2.**

$$(1 + \rho^2 - 2\rho x)^2 \operatorname{tr} \left( \mathbf{T}_m(x) \mathbf{R}_m^{-1} \right) = (m+1)(1 + \rho^2 - 2\rho x) + 1 - \rho^2 - 2 \frac{1 - \rho^{m+1} T_{m+1}(x)}{\det \mathbf{R}_m}. \quad (26)$$

*Proof.* Denote  $r = \frac{1}{2} \left( \rho + \frac{1}{\rho} \right)$ . Then,  $1 + \rho^2 - 2\rho x = -2\rho(x - r)$ .

From Lemma 3.1 it follows that (26) is equivalent to

$$4\rho^2(x - r)^2 \operatorname{tr} \left( \mathbf{T}_m(x) \mathbf{R}_m^{-1} \right) = \frac{2\rho}{U_m(r)} (T_{m+1}(x) - T_{m+1}(r)) - 2\rho(m+1)(x - r). \quad (27)$$



We see that  $\mathbf{R}_m^{-1} = \frac{1}{\rho} \mathbf{A}$ , where  $\mathbf{A}$  is a symmetric matrix with entries defined by (18). Note that the symmetric Toeplitz structure of the matrix  $\mathbf{T}_m$  and the fact that  $\mathbf{A}$  is symmetric imply

$$\rho \operatorname{tr} \left( \mathbf{T}_m(x) \mathbf{R}_m^{-1} \right) = 2 \sum_{k=0}^{m-1} T_k(x) \sum_{i=1}^{m-k} a_{i,i+k}.$$

We interchange two sums in the above equation. Then we have

$$\rho \operatorname{tr} \left( \mathbf{T}_m(x) \mathbf{R}_m^{-1} \right) = 2 \sum_{i=1}^m \sum_{k=0}^{m-i} T_k(x) a_{i,i+k}$$

From (18) we get

$$(\rho U_m(r)) \operatorname{tr} \left( \mathbf{T}_m(x) \mathbf{R}_m^{-1} \right) = 2 \sum_{i=1}^m U_{i-1}(r) \sum_{k=0}^{m-i} T_k(x) U_{m-i-k}(r). \tag{28}$$

From (20) it follows that

$$2(x-r) \sum_{k=0}^{m-i} T_k(x) U_{m-i-k}(r) = T_{m-i+1}(x) - T_{m-i+1}(r). \tag{29}$$

This together with (28) gives

$$(x-r)(\rho U_m(r)) \operatorname{tr} \left( \mathbf{T}_m(x) \mathbf{R}_m^{-1} \right) = \sum_{i=1}^m U_{i-1}(r) (T_{m-i+1}(x) - T_{m-i+1}(r)),$$

which can be rewritten as follows

$$(x-r)(\rho U_m(r)) \operatorname{tr} \left( \mathbf{T}_m(x) \mathbf{R}_m^{-1} \right) = S_1 - S_2, \tag{30}$$

where

$$S_1 = \sum_{i=1}^m U_{i-1}(r) T_{m-i+1}(x), \quad S_2 = \sum_{i=1}^m U_{i-1}(r) T_{m-i+1}(r). \tag{31}$$

From (19) we have

$$S_2 = \sum_{j=1}^m T_j(r) U_{m-j}(r) = \frac{m}{2} U_m(r).$$

Note that (20) implies

$$2(x-r)S_1 = 2(x-r) \sum_{j=1}^m T_j(x) U_{m-j}(x) = (T_{m+1}(x) - T_{m+1}(r)) - (x-r)U_m(r).$$

This together with (30) gives

$$4(x-r)^2(\rho U_m(r)) \operatorname{tr} \left( \mathbf{T}_m(x) \mathbf{R}_m^{-1} \right) = 4(x-r)(S_1 - S_2).$$

Finally, we obtain

$$4(x-r)^2(\rho U_m(r)) \operatorname{tr} \left( \mathbf{T}_m(x) \mathbf{R}_m^{-1} \right) = 2(T_{m+1}(x) - T_{m+1}(r)) - 2(m+1)(x-r)U_m(r). \tag{32}$$

From this (27) follows immediately. The proof of (26) is now complete.  $\square$

## 6. Conclusions

This paper shows, through a particular example, why sampling survey methodology needs mathematics and vice versa, how it can be a source of intriguing purely mathematical problems. We were concerned with a connection between rotation sampling design and the Chebyshev polynomials, which was used in KW to give a complete description of the recursion for BLUEs of means on every occasion. The recursion depth was identified through the largest gap in the rotation pattern and the recursion coefficients in terms of the Chebyshev polynomials depending on correlations for a single unit. According to the standard Patterson model, these correlations are assumed to be exponential in time and the same for every unit, with independence between units. The general form of the recursion was derived in KW under ASSUMPTIONS I and II and expressed in terms of the Chebyshev polynomials. There is a strong numerical evidence that both the assumptions are not needed for the recursion to hold true. In this paper, using intrinsic properties of the Chebyshev polynomials of the first and the second kind, we proved that, at least for rotation designs with one arbitrary large gap, ASSUMPTION I is always satisfied. However, the problem if ASSUMPTION II is also satisfied, even in such a simplified rotation pattern, remains a challenging mathematical question.

## References

- Bell, P., (2001). Comparison of alternative Labour Force Survey estimators. *Survey Methodology*, 27, pp. 53–63.
- Cantwell, P. J., (1990). Variance formulae for the composite estimators in rotation designs. *Survey Methodology*, 16, pp. 153–163.
- Cantwell, P. J., Caldwell, C. V., (1998). Examining the revisions in monthly retail and wholesale trade surveys under rotation panel design. *Journal of Official Statistics* 14, pp. 47–54.
- Ciepiela, P., Gniado, M., Wesołowski, J., Wojtyś, M., (2012). Dynamic  $K$ -composite estimator for an arbitrary rotation scheme. *Statistics in Transition*, 13(1), pp. 7–20.
- Fuller, W. A., Rao, J. N. K., (2001). A regression composite estimator with application to the Canadian Labour Force Survey. *Survey Methodology*, 27, pp. 45–51.
- Gurney, M., Daly, J. F., (1965A). multivariate approach to estimation in periodic sample surveys. *Proceedings of the Survey Research Method Section American Statistical Association*, pp. 242–257.
- Hansen, M. H., Hurwitz, W.N., Nisselson, H., Steinberg, J., (1955). The redesign of the census current population survey. *Journal of the American Statistical Association*, 50, pp. 701–719.
- Kalton, G., (2023). Probability vs. nonprobability sampling: from the birth of survey sampling to the present day. *Statistics in Transition*.
- Karna, J. P., Nath, D. C., (2015). Rotation sampling: introduction and review of recent developments. *Journal of Assam Science Society*, 56(2), pp. 90–111.

- Kordos, J., (2012). Application of rotation methods in sample surveys in Poland. *Statistics in Transition*, 13(2), pp. 243–260.
- Kowalczyk, B., Juszczyk, D., (2018). Composite estimator based on the recursive ratio for an arbitrary rotation scheme. *Mathematical Population Studies*, 25(4), pp. 227–247.
- Kowalski J., (2009). Optimal estimation in rotation patterns. *Journal of Statistical Planning and Inference*, 139, pp. 1405–1420.
- Kowalski, J., Wesołowski, J., (2015). Exploring recursion for optimal estimators under cascade rotation. *Survey Methodology*, 41(1), pp. 99–126.
- McLaren, C., Steel, D., (2000). The impact of different rotation patterns on sampling variance of seasonally adjusted and trend estimates. *Survey Methodology*, 26, pp. 163–172.
- Paszkowski, S., (1975). *Numerical applications of Chebyshev polynomials*, Warsaw (in Polish).
- Patterson, H. D., (1950). Sampling on successive occasions. *Journal of the Royal Statistical Society, Ser. B*, 12, pp. 241–255.
- Rao, J. N. K., Graham, J. E., (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, pp. 492–509.
- Särndal, C.-E., Swensson, B., Wretman, J., (1992). *Model Assisted Survey Sampling*, Springer.
- Singh, A. C., Kennedy, B., Wu, S., (2001). Regression composite estimation for the Canadian Labour Force Survey with a rotating panel design. *Survey Methodology*, 27, pp. 33–44.
- Steel, D., McLaren, C., (2002). In search of a good rotation pattern. In: *Advances in Statistics, Combinatorics and Related Areas*, World Scientific, pp. 309–319.
- Steel, D., McLaren, C., (2008). Design and analysis of repeated surveys. *Working Paper*, Center for Statist. Survey Meth., Wollonong Univ., 11-08, pp. 1–13.
- Szegő, G., (1959). *Orthogonal polynomials*, New York.
- Wesołowski, J., (2010). Recursive optimal estimation in Szarkowski rotation scheme. *Statistics in Transition*, 11(2), pp. 267–285.
- Yansaneh, I. S., Fuller, W., (1998). Optimal recursive estimation for repeated surveys. *Survey Methodology* 24, pp. 31–40.



# A nonparametric analysis of discrete time competing risks data: a comparison of the cause-specific-hazards approach and the vertical approach

Bonginkosi Duncan Ndlovu<sup>1</sup>, Sileshi Fanta Melesse<sup>2</sup>,  
Temesgen Zewotir<sup>3</sup>

## ABSTRACT

Nicolaie et al. (2010) have advanced a vertical model as the latest continuous time competing risks model. The main objective of this article is to re-cast this model as a nonparametric model for analysis of discrete time competing risks data. Davis and Lawrance (1989) have advanced a cause-specific-hazard driven method for summarizing discrete time data non-parametrically. The secondary objective of this article is to compare the proposed model to this model. We pay particular attention to the estimates for the cause-specific-hazards and the cumulative incidence functions as well as their respective standard errors.

**Key words:** vertical model; total hazards; relative hazards; cause-specific-hazards; cumulative incidence functions.

## 1. The first section

Competing risks have come to refer to survival analysis experiments where subjects may fail by more than one mode of failure. The vertical model (Nicolaie et al., 2010) is the latest competing risks model that has been advanced additional to the models proposed by Prentice et al.(1978) and Larson and Dinse (1985). The model proposes total hazards and relative hazards for modelling competing risks data. It is the only competing risks model that is capable of handling the standard competing risks data as well as data that has unknown failure causes for some subjects, that is, the model is invariant to the presence or absence of unknown failure causes (Nicolaie et al., 2015). Furthermore, Nicolaie et al. (2018) have extended the model for analysis of data that has a sizable proportion of cured subjects. Both these topics, i.e., handling data with missing failure causes and data that comes with cured subjects, have not received satisfactory attention in discrete time. Whilst the vertical model possesses these attractive features, it cannot be naively applied to discrete time competing risks data because the model was introduced as a continuous time model. The main objective of this article is to modify this model and present it as a nonparametric discrete time competing risks model additional to the nonparametric model suggested by Davis and Lawrance (1989). The complication with discrete time data is excessive number of ties. Continuous time competing risks models are premised on the factorization of

<sup>1</sup>Department of Statistics, Durban University of Technology, South Africa. E-mail: bongi@dut.ac.za

<sup>2</sup>Department of Statistics, University of KwaZulu-Natal, South Africa. E-mail: melesse@ukzn.ac.za

<sup>3</sup>Department of Statistics, University of KwaZulu-Natal, South Africa. E-mail: zewotir@ukzn.ac.za

ORCID: <https://orcid.org/0000-0003-1438-8571>.

© B. D. Ndlovu, S. F. Melese, T. Zewotir. Article available under the CC BY-SA 4.0 licence



the full likelihood function into cause-specific likelihood functions assumption. This is not possible in the presence of a disproportionately large number of ties. The model suggested by Davis and Lawrance (1989) is one the models that have been advanced specifically for analysis of discrete time competing risks data. In fact, this model is the first truly discrete time competing risks model to be advanced in the competing risks literature. This model was followed by the multinomial model (Ambrogi et al., 2009, Tutz and Schmid, 2016) as the first regression model for analysis of discrete time data. Here, data are modelled with discrete time version of cause-specific-hazards. Concerns have been expressed about this model owing to estimation of a significantly larger number of cause-specific-hazard parameters simultaneously. Recently, Lee et al. (2018) have advanced an alternate regression model, which addresses these reservation regarding the multinomial model (Ambrogi et al., 2009, Tutz and Schmid, 2016) where the cause-specific-hazards are estimated individually via the application of a binomial distribution within the GEE framework. Both these models, that is, the multinomial model (Ambrogi et al., 2009, Tutz and Schmid, 2016) and the binomial model (Lee et al., 2018) give rise to a regression model for the cumulative incidence function, which has become notorious for complicating the assessment of covariate effects. Berger et al. (2020) have since proposed a discrete time subhazard regression model for the cumulative incidence function to address the limitations of the cause-specific-hazard denominated regression model for the cumulative incidence function. The Davis and Lawrance (1989) model proposes nonparametric discrete time cause-specific-hazards for modelling data. We shall refer to this model as the cause-specific-hazards model. If the vertical model proposes total hazards and relative hazards for characterizing data, it implies that the standard summary statistics are now obtained from the total hazard and relative hazard estimates in the place of the more familiar cause-specific-hazard estimates as suggested by the cause-specific-hazards model. The most logical question that follows is whether the two estimation methods produce the same estimates for the same quantity. For example, if one method proposes direct estimation of cause-specific-hazards from data and the other one suggests that the estimates for the same quantities are now derived from total and relative hazard estimates, then how do the two methods compare, and more importantly, how do the standard errors for the two estimation methods compare. As a secondary objective of this article, we attempt to address these questions. With the cause-specific-hazards and the cumulative incidence functions as the most widely quoted pair for summarizing competing risks data, we will focus on these quantities and their respective standard errors to address these pertinent questions.

As already highlighted, Davis and Lawrance (1989) proposes cause-specific-hazards for modelling data. Let  $\tilde{T}$  and  $D$  represent time to failure and failure type respectively, where  $D \in \{1, 2, \dots, J\}$  and  $J$  is the number of failure causes. Also, let  $C$  denote the time to censoring. Observed competing risks data can be represented by  $y_i = (t_i, D_i)$  for  $i = 1, \dots, n$  where  $T_i = \min\{\tilde{T}_i, C_i\}$ , such that  $T_i = t_i$  is a failure time due to failure type  $j$  or censoring time according to whether  $D_i = j$  or 0. It is assumed that time to failure and censoring time are in discrete units, i.e.,  $\tilde{T}, C \in \{1, \dots, q\}$  where  $q$  is a positive integer. The definition of discrete time cause-specific-hazards for nonparametric purposes is given by

$$\lambda_j(t) = P(T = t; D = j | T \geq t) \quad (1)$$

for  $t = 1, 2, \dots, q$  and  $j = 1, 2, \dots, J$ . Suppose that at time  $t$ ,  $n_{(t)}$  and  $d_{(jt)}$  denote the number at risks and the number of failures due to failure type  $j$ , respectively. Davis and Lawrance (1989) have shown that the observed data likelihood function is a kernel of a multinomial likelihood function:

$$\mathcal{L} = \sum_{s=1}^q \sum_{j=1}^J d_{(js)} \log \lambda_j(s) + (n_{(s)} - d_{(s)}) \log(1 - \lambda(s)) \tag{2}$$

where  $d_{(s)} = \sum_{j=1}^J d_{(js)}$  and  $\lambda(s) = \sum_{j=1}^J \lambda_j(s)$ . As such, the MLE for  $\lambda_j(s)$  is given by

$$\hat{\lambda}_j(s) = d_{(js)} / n_{(s)}$$

for  $s = 1, 2, \dots, q$  and  $j = 1, 2, \dots, J$ . The estimates for the cumulative incidence functions are then given by

$$\hat{F}_j(t) = \sum_{s=1}^t \hat{S}(s-1) \hat{\lambda}_j(s) \tag{3}$$

for  $t = 1, 2, \dots, q$  and  $j = 1, 2, \dots, J$ , where  $\hat{S}(t) = \prod_{s=1}^t (1 - \hat{\lambda}(s))$ .

The vertical model proposes a factorization of the bivariate distribution of failure time and failure type into a marginal distribution for failure time and a distribution for failure type conditional on failure time as characterized via total hazards and failure type probabilities conditional on failure time (relative hazards), respectively. For analysis of discrete time competing risks data nonparametrically, we propose the following definition for discrete time total hazards:

$$\lambda(t) = P(T = t | T \geq t) = \sum_{j=1}^J P(T = t; D = j | T \geq t) = \sum_{j=1}^J \lambda_j(t)$$

for  $t = 1, 2, \dots, q$ . The total hazard  $\lambda(t)$  is the probability of failure, by any cause, at time  $t$  given survival to time  $t$ . On the other hand, the relative hazard  $\pi_j(t)$  is the probability that a failure is attributable to cause  $j$  given that a failure has occurred at time  $t$ . The definition of relative hazards is given by

$$\pi_j(t) = P(D = j | T = t)$$

$t = 1, 2, \dots, q$  and for  $j = 1, 2, \dots, J$ . The term "relative hazards" comes from:

$$\begin{aligned} \pi_j(t) &= P(D = j | T = t) = \frac{P(D = j, T = t)}{P(T = t)} = \frac{P(D = j, T = t, T \geq t) / P(T \geq t)}{P(T = t, T \geq t) / P(T \geq t)} \\ &= \frac{\lambda_j(t)}{\sum_{j=1}^J \lambda_j(t)} \end{aligned} \tag{4}$$

It follows from (4) that:

$$\lambda_j(t) = \lambda(t) \pi_j(t) \tag{5}$$

Thus, the cause-specific-hazard estimates are now estimated indirectly from total hazard and relative hazard estimates via (5). All failures contribute to the estimation of the total hazards, then, the total hazards are apportioned to cause-specific-hazards via relative hazards. This formulation become very convenient in the presence of subjects with missing failure causes because these subjects also contribute to the estimation of total hazards. The expression for the cumulative incidence function is also given in terms of total and relative hazards:

$$F_j(t) = \sum_{s=1}^t S(s-1)\lambda(s)\pi_j(s)$$

$t = 1, 2, \dots, q$  and for  $j = 1, 2, \dots, J$ .

This concludes the exercise of re-framing the vertical model as a discrete time nonparametric competing risks model. To determine the summary statistics, i.e., the estimates for cause-specific-hazards and cumulative incidence functions we require the estimates for total hazards and relative hazards. Let  $\theta = (\pi^T, \lambda^T)^T$  where  $\lambda = (\lambda(1), \lambda(2) \dots \lambda(q))^T$ ,  $\pi = (\pi_1^T, \pi_2^T \dots \pi_{j-1}^T)^T$ , and  $\pi_j = (\pi_j(1), \pi_j(2) \dots \pi_j(q))^T$ . In Section 2 we demonstrate the estimation of total hazards and relative hazards, that is, we determine  $\hat{\theta}$ . This is followed by the application of the proposed model in Section 3. We derive the standard errors for the cumulative incidence function estimates in Section 4. This concludes the first part of our twofold objectives. In Section 5, we address the second part of our objective, that is, to prove that the estimates for cause-specific-hazards and cumulative incidence function as well as the corresponding standard errors are identical by the proposed model or the cause-specific-hazards model. We conclude the article with a discussion in Section 6.

## 2. Estimation

It is straightforward to determine the MLE's for the total hazards and relative hazards. The observed data likelihood function which is specified in terms of total hazards and relative hazards is differentiated with respect to these quantities. When the vertical model is assumed,  $P(T_i = t_i, D_i = j)$ , the contribution of subject  $i$  that failed at time  $t_i$  due to failure cause  $j$  to the observed data likelihood function is now replaced by  $P(D_i = j | T_i = t_i)P(T_i = t_i)$  while a censored subject  $i$  continues to contribute  $P(T_i > t_i)$ . Define an indicator variable  $d_{ij}$  such that  $d_{ij}$  is 1 or 0 according to whether subject  $i$  failed by cause  $j$  or not and let  $d_i = \sum_{j=1}^J d_{ij}$  where  $d_i$  indicates failure by any cause for subject  $i$ . The observed data log-likelihood function can be written as:

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_{i=1}^n \sum_{j=1}^J d_{ij} \log P(D_i = j | T_i = t_i) P(T_i = t_i) + (1 - d_i) \log P(T_i > t_i) \\ &= \sum_{i=1}^n \sum_{j=1}^J d_{ij} \log \pi_j(t_i) + \sum_{i=1}^n d_i \log P(T_i = t_i) + (1 - d_i) \log P(T_i > t_i) \\ &= \mathcal{L}(\pi) + \mathcal{L}(\lambda) \end{aligned}$$



We can ignore  $\mathcal{L}(\pi)$  because the estimates for the relative hazards can be obtained from (4), that is:

$$\hat{\pi}_j(t) = \frac{\hat{\lambda}_j(t)}{\sum_{j=1}^J \hat{\lambda}_j(t)} = \frac{d_{(jt)}/n_{(t)}}{\sum_{j=1}^J d_{(jt)}/n_{(t)}} = \frac{d_{(jt)}}{d_{(t)}}$$

The log-likelihood function  $\mathcal{L}(\lambda)$  is a failure time log-likelihood function. It is straightforward to show that  $\mathcal{L}(\lambda)$  can be written as:

$$\mathcal{L}(\lambda) = \sum_{s=1}^q d_{(s)} \log \lambda(s) + (n_{(s)} - d_{(s)}) \log(1 - \lambda(s))$$

Naturally,  $\frac{\partial \mathcal{L}(\lambda)}{\partial \lambda(s)} = 0$  yields an MLE for  $\lambda_j(s)$  given by

$$\hat{\lambda}(s) = \frac{d_{(s)}}{n_{(s)}}$$

The estimates for total hazards and relative hazards can be plugged into appropriate equations to recover the estimates for cause-specific-hazards and cumulative incidence functions. In the next section we demonstrate the application of the proposed model.

### 3. Application

We apply the proposed model to data that comes with Ecdat R package (Croissant and Graves, 2020). In these data 3343 recently unemployed individuals are tracked the moment they lost their jobs until they are re-employed into part-time employment (339), full-time employment (1073) or are censored (1255). Of the remaining 676, 574 were re-employed but the type employment was not recorded. It is not clear with the other 102 subjects if they were censored or were re-employed. We have excluded the 674 individuals to leave a final sample of 2667 that were considered for analysis. Failure times assume values in  $\{1, 2, \dots, 26, 27, 28\}$  where time is measured in bi-weekly units. There are some covariates that come with data such as unemployment benefits, disregard rate, replacement rate, etc., which are naturally ignored.

In the application of the proposed model we have computed the relative hazard and total hazard estimates, respectively from:

$$\hat{\pi}_j(t) = \frac{d_{(jt)}}{d_{(t)}} \text{ and } \hat{\lambda}(t) = \frac{d_{(t)}}{n_{(t)}}$$

The variances are respectively given by

$$V(\hat{\pi}_j(t)) = \frac{\hat{\pi}_j(t)(1 - \hat{\pi}_j(t))}{d_{(t)}} \text{ and } V(\hat{\lambda}(t)) = \frac{\hat{\lambda}(t)(1 - \hat{\lambda}(t))}{n_{(t)}}$$

These estimates are listed in Table 1 together with corresponding standard errors. We have labelled full-time re-employment as cause 1 and part-time re-employment as cause 2.

Table 1: Maximum likelihood estimates for the total and relative hazards from the Vertical Model as well as the cause-specific-hazards estimates from the Cause-Specific-Hazards Model (with standard errors)

	<b>Model I</b>		<b>Model II</b>	
	Nonparametric Vertical Model		Cause-Specific-Hazards Model	
	$\hat{\pi}_1$	$\hat{\lambda}$	$\hat{\lambda}_1$	$\hat{\lambda}_2$
T1	0.752(0.022)	0.147(0.007)	0.110(0.006)	0.037(0.004)
T2	0.761(0.028)	0.104(0.006)	0.079(0.006)	0.025(0.003)
T3	0.763(0.034)	0.081(0.006)	0.062(0.006)	0.019(0.003)
T4	0.727(0.051)	0.048(0.005)	0.035(0.005)	0.013(0.003)
T5	0.748(0.037)	0.098(0.008)	0.074(0.006)	0.025(0.004)
T6	0.762(0.066)	0.037(0.006)	0.028(0.005)	0.009(0.003)
T7	0.779(0.039)	0.107(0.009)	0.083(0.009)	0.024(0.005)
T8	0.625(0.098)	0.029(0.006)	0.019(0.005)	0.011(0.004)
T9	0.825(0.060)	0.055(0.008)	0.045(0.008)	0.001(0.004)
T10	0.060(0.204)	0.009(0.004)	0.005(0.003)	0.005(0.003)
T11	0.838(0.066)	0.054(0.009)	0.046(0.009)	0.009(0.004)
T12	0.700(0.145)	0.021(0.007)	0.015(0.005)	0.006(0.004)
T13	0.756(0.075)	0.075(0.013)	0.057(0.011)	0.018(0.006)
T14	0.833(0.062)	0.099(0.016)	0.083(0.014)	0.017(0.007)
T15	0.864(0.073)	0.081(0.017)	0.069(0.015)	0.011(0.006)
T16	0.769(0.117)	0.059(0.016)	0.046(0.014)	0.014(0.008)
T17	0.889(0.105)	0.050(0.016)	0.044(0.015)	0.006(0.006)
T18	0.778(0.139)	0.059(0.019)	0.046(0.017)	0.013(0.009)
T19	0.667(0.192)	0.044(0.018)	0.029(0.015)	0.015(0.010)
T20	1.000(0.000)	0.025(0.014)	0.025(0.014)	0.000(0.000)
T21	0.571(0.187)	0.071(0.026)	0.041(0.019)	0.030(0.017)
T22	0.800(0.179)	0.067(0.029)	0.053(0.026)	0.013(0.013)
T23	0.000(0.000)	0.016(0.016)	0.000(0.000)	0.016(0.016)
T24	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)
T25	0.000(0.000)	0.019(0.018)	0.000(0.000)	0.019(0.018)
T26	1.000(0.000)	0.045(0.031)	0.045(0.031)	0.000(0.000)
T27	0.833(0.152)	2.000(0.073)	0.167(0.068)	0.033(0.033)

Since  $\hat{\pi}_1(t) + \hat{\pi}_2(t) = 1$ , we have only listed  $\hat{\pi}_1(t)$ . We have also listed the cumulative incidence function estimates together with corresponding standard errors in Table 2.

The cumulative incidence function estimates are obtained from:

$$\hat{F}_j(t) = \sum_{s=1}^t \hat{S}(s-1) \hat{\lambda}(s) \hat{\pi}_j(s)$$

Table 2: Maximum likelihood estimates for the Cumulative Incidence Function from the Vertical Model (with standard errors)

Nonparametric Vertical Model		
	$\hat{F}_1$	$\hat{F}_2$
T1	0.110(0.006)	0.036(0.004)
T2	0.178(0.007)	0.058(0.005)
T3	0.225(0.008)	0.072(0.005)
T4	0.249(0.009)	0.082(0.005)
T5	0.299(0.009)	0.098(0.006)
T6	0.316(0.009)	0.103(0.006)
T7	0.364(0.010)	0.117(0.007)
T8	0.374(0.010)	0.123(0.007)
T9	0.397(0.011)	0.128(0.007)
T10	0.399(0.011)	0.130(0.007)
T11	0.421(0.011)	0.134(0.007)
T12	0.427(0.011)	0.137(0.008)
T13	0.452(0.012)	0.145(0.008)
T14	0.485(0.012)	0.152(0.008)
T15	0.511(0.013)	0.156(0.009)
T16	0.526(0.013)	0.162(0.009)
T17	0.539(0.014)	0.162(0.009)
T18	0.553(0.014)	0.166(0.009)
T19	0.562(0.015)	0.169(0.009)
T20	0.569(0.015)	0.169(0.009)
T21	0.579(0.016)	0.178(0.011)
T22	0.592(0.016)	0.181(0.011)
T23	0.592(0.016)	0.185(0.012)
T24	0.000(0.000)	0.000(0.000)
T25	0.592(0.016)	0.189(0.012)
T26	0.602(0.017)	0.189(0.012)
T27	0.637(0.021)	0.196(0.014)

The standard errors for the cumulative incidence function estimates as derived in the next section are given by

$$V(\hat{F}_j(t)) = \sum_{s=1}^t \text{Var}(\hat{S}(s-1)\hat{\lambda}(s)\hat{\pi}_j(s)) + 2 \sum_{s=1}^{t-1} \sum_{k=s+1}^t \text{Cov}(\hat{S}(s-1)\hat{\lambda}(s)\hat{\pi}_j(s), \hat{S}(k-1)\hat{\lambda}(k)\hat{\pi}_j(k))$$

where:

$$V(\hat{S}(s-1)\hat{\lambda}(s)\hat{\pi}_j(s)) = (\hat{S}(s-1)\hat{\lambda}_j(s)\hat{\pi}_j(s))^2 \left( \sum_{l=1}^{s-1} \frac{d_{(l)}}{n_{(l)}(n_{(l)} - d_{(l)})} + \frac{n_{(s)} - d_{(s)}}{d_{(s)}n_{(s)}} + \frac{d_{(s)} - d_{(js)}}{d_{(s)}d_{(js)}} \right)$$

and,

$$\begin{aligned} \text{Cov}(\hat{S}(s-1)\hat{\lambda}(s)\hat{\pi}_j(s)\hat{S}(k-1)\hat{\lambda}(k)\hat{\pi}_j(k)) &= (\hat{S}(s-1)\hat{\lambda}(s)\hat{\pi}_j(s)\hat{S}(k-1)\hat{\lambda}(k)\hat{\pi}_j(k)) \\ &\times \left( \sum_{l=1}^{s-1} \frac{d_{(l)}}{n_{(l)}(n_{(l)} - d_{(l)})} - \frac{1}{n_{(s)}} \right) \end{aligned}$$

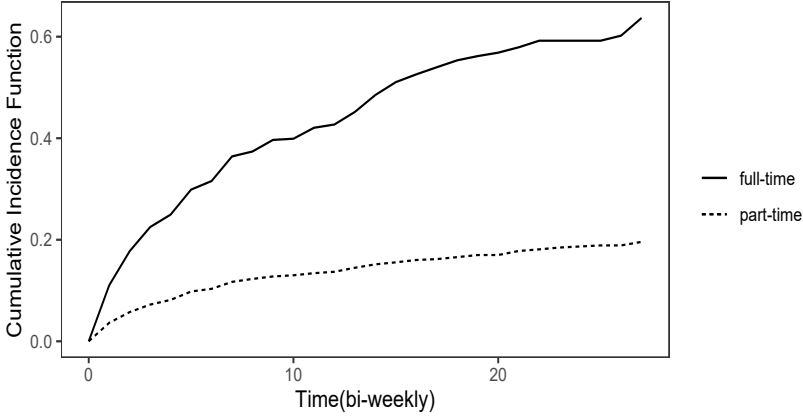


Figure 1: The cumulative incidence function of exit to full-time and part-time

In Figure 1 we have plotted the cumulative incidence function estimates from the proposed model because the proposed model and the cause-specific-hazards model produce identical estimates for cumulative incidence function as will be shown in Section 5. Clearly, the plot suggest that unemployed subjects are more likely to exit the state of unemployment to full-time employment than than to part-time employment.

#### 4. Cumulative Incidence Function Standard Errors

The expression of standard errors for the cumulative incidence function estimates under the vertical model is given by

$$\begin{aligned} V(\hat{F}_j(t)) &= \sum_{s=1}^t \text{Var}(S(s-1)\lambda(s)\pi_j(s)) \\ &+ 2 \sum_{s=1}^{t-1} \sum_{k=s+1}^t \text{Cov}(S(s-1)\lambda(s)\pi_j(s), S(k-1)\lambda(k)\pi_j(k)) \Big|_{\theta=\hat{\theta}} \end{aligned}$$

Let  $Q_s(\lambda_1, \dots, \lambda_{s-1}, \lambda_s, \pi_j(s)) = S(s-1)\lambda(s)\pi_j(s)$  and  $Q_k(\lambda_1, \dots, \lambda_{s-1}, \lambda_s, \dots, \lambda_{k-1}, \lambda_k, \pi_j(k)) = S(k-1)\lambda(k)\pi_j(k)$ , where  $s < k$ . We begin by com-

putting the partial derivatives:

$$\begin{aligned} \frac{\partial Q_s}{\partial S(s-1)} &= \lambda(s)\pi_j(s) \\ \frac{\partial Q_s}{\partial \lambda(s)} &= S(s-1)\pi_j(s) \\ \frac{\partial Q_s}{\partial \pi_j(s)} &= S(s-1)\lambda(s) \\ \frac{\partial Q_k}{\partial S(s-1)} &= \frac{Q_k}{S(s-1)} \\ \frac{\partial Q_k}{\partial \lambda(s)} &= -\frac{Q_k}{1-\lambda(s)} \end{aligned}$$

Assuming that  $d_{(1)}, d_{(2)} \dots d_{(s)}$  are uncorrelated (Dinse and Larson, 1986), then  $\text{Cov}(\lambda(l), \lambda(m)) = 0$  when  $l \neq m$  for  $l = 1, 2 \dots q$  and  $m = 1, 2 \dots q$ . It, therefore, follows that:

$$\begin{aligned} V(Q_s) &= \begin{pmatrix} \lambda(s)\pi_j(s) \\ S(s-1)\pi_j(s) \\ S(s-1)\lambda(s) \end{pmatrix} \begin{pmatrix} V(S(s-1)) & 0 & 0 \\ 0 & V(\lambda(s)) & 0 \\ 0 & 0 & V(\pi_j(s)) \end{pmatrix} \begin{pmatrix} \lambda(s)\pi_j(s) \\ S(s-1)\pi_j(s) \\ S(s-1)\lambda(s) \end{pmatrix} \\ &= (\lambda(s)\pi_j(s))^2 \text{Var}(S(s-1)) + (S(s-1)\pi_j(s))^2 V(\lambda(s)) \\ &\quad + (S(s-1)\lambda(s))^2 V(\pi_j(s)) \\ &= (S(s-1)\lambda(s)\pi_j(s))^2 \sum_{l=1}^{s-1} \frac{\lambda(l)}{n_{(l)}(1-\lambda(l))} + \frac{1-\lambda(s)}{\lambda(s)n_{(s)}} + \frac{1-\pi_j(s)}{d_{(s)}\pi_j(s)} \end{aligned}$$

Thus,

$$\begin{aligned} V(\hat{S}(s-1)\hat{\lambda}(s)\hat{\pi}_j(s)) &= (S(s-1)\lambda(s)\pi_j(s))^2 \sum_{l=1}^{s-1} \frac{\lambda(l)}{n_{(l)}(1-\lambda(l))} + \frac{1-\lambda(s)}{\lambda(s)n_{(s)}} \\ &\quad + \frac{1-\pi_j(s)}{d_{(s)}\pi_j(s)} \Big|_{\theta=\hat{\theta}} \\ &= (\hat{S}(s-1)\hat{\lambda}_j(s)\hat{\pi}_j(s))^2 \left( \sum_{l=1}^{s-1} \frac{d_{(l)}}{n_{(l)}(n_{(l)}-d_{(l)})} \right. \\ &\quad \left. + \frac{n_{(s)}-d_{(s)}}{d_{(s)}n_{(s)}} + \frac{d_{(s)}-d_{(js)}}{d_{(s)}d_{(js)}} \right) \end{aligned}$$

We now consider:

$$\begin{aligned} \text{Cov}(Q_s, Q_k) &= (S(s-1))^2 \lambda(s) \pi_j(s) \frac{S(k-1) \lambda(k) \pi_j(k)}{S(s-1)} \frac{\sum_{l=1}^{s-1} \lambda(l) (1 - \lambda(l))}{n(l)} \\ &\quad - S(s-1) \pi_j(s) \frac{S(k-1) \lambda(k) \pi_j(k)}{1 - \lambda(s)} \frac{\lambda(s) (1 - \lambda(s))}{n(s)} \\ &= S(s-1) \lambda(s) \pi_j(s) S(k-1) \lambda(k) \pi_j(k) \left( \sum_{l=1}^{s-1} \frac{\lambda(l) (1 - \lambda(l))}{n(l)} - \frac{1}{n(s)} \right) \end{aligned}$$

Thus,

$$\begin{aligned} \text{Cov}(\hat{S}(s-1) \hat{\lambda}(s) \hat{\pi}_j(s) \hat{S}(k-1) \hat{\lambda}(k) \hat{\pi}_j(k)) &= \text{Cov}(Q_s, Q_k) \Big|_{\theta=\hat{\theta}} \\ &= (\hat{S}(s-1) \hat{\lambda}(s) \hat{\pi}_j(s) \hat{S}(k-1) \hat{\lambda}(k) \hat{\pi}_j(k)) \\ &\quad \times \left( \sum_{l=1}^{s-1} \frac{d(l)}{n(l)(n(l) - d(l))} - \frac{1}{n(s)} \right) \end{aligned}$$

## 5. Proofs

In this section we demonstrate that estimates for cause-specific-hazards and cumulative incidence function together with corresponding standard errors derived from the proposed model and the cause-specific-hazards model are identical. Let  $\hat{\lambda}_j^V(t)$  and  $\hat{\lambda}_j^C(t)$  denote the estimates for the cause-specific-hazards via the proposed model and the cause-specific-hazards model, respectively. Likewise, let  $\hat{F}_j^V(t)$  and  $\hat{F}_j^C(t)$  represent the estimates for the cumulative incidence function that are produced by the proposed model and the cause-specific-hazards model, respectively.

Beginning with the estimates for the cause-specific-hazards:

$$\hat{\lambda}_j^V(t) = \hat{\pi}_j(t) \hat{\lambda}(t) = \frac{d(jt)}{d(t)} \times \frac{d(t)}{n(t)} = \frac{d(jt)}{n(t)} = \hat{\lambda}_j^C(t)$$

It follows that the cumulative incidence function estimates by both models are identical, that is:

$$\hat{F}_j^V(t) = \sum_{s=1}^t \hat{S}(s-1) \hat{\lambda}(s) \hat{\pi}_j(s) = \sum_{s=1}^t \hat{S}(s-1) \hat{\lambda}_j^C(s) = \hat{F}_j^C(t)$$

To determine the standard errors for cause-specific-hazard and cumulative incidence function estimates, we apply the delta method. We begin with standard errors for the cause-

specific-hazard estimates. The standard error for  $\hat{\lambda}_j^C(s) = \frac{d(s)}{n(s)}$  is well known and it is given by

$$V(\hat{\lambda}_j^C(s)) = \frac{\hat{\lambda}_j(s)(1 - \hat{\lambda}_j(s))}{n(s)}$$

We now determine the expression for the variance of  $\hat{\lambda}_j^V(s) = \hat{\lambda}(s)\hat{\pi}_j(s)$ . Since

$$\frac{\partial \mathcal{L}}{\partial \lambda(l) \partial \pi_j(m)} = 0$$

for  $l = 1, \dots, q; j = 1, 2, \dots, J; m = 1, \dots, q$ , thus:

$$\begin{aligned} V(\hat{\lambda}_j^V(s)) &= \begin{pmatrix} \frac{\partial \lambda_j^V(s)}{\partial \lambda(s)} & \frac{\partial \lambda_j^V(s)}{\partial \pi_j(s)} \end{pmatrix} \begin{pmatrix} V(\lambda(s)) & 0 \\ 0 & V(\pi_j(s)) \end{pmatrix} \\ &\times \begin{pmatrix} \frac{\partial \lambda_j^V(s)}{\partial \lambda(s)} & \frac{\partial \lambda_j^V(s)}{\partial \pi_j(s)} \end{pmatrix}^T \Big|_{\theta = \hat{\theta}} \\ &= \pi_j(s)^2 V(\lambda(s)) + \lambda(s)^2 V(\pi_j(s)) \Big|_{\theta = \hat{\theta}} \\ &= \hat{\pi}_j(s)^2 V(\hat{\lambda}(s)) + \hat{\lambda}(s)^2 V(\hat{\pi}_j(s)) \end{aligned}$$

where the partial derivatives are given by

$$\begin{aligned} \frac{\partial \lambda_j^V(s)}{\partial \lambda(s)} &= \pi_j(s) \\ \frac{\partial \lambda_j^V(s)}{\partial \pi_j(s)} &= \lambda(s) \end{aligned}$$

Therefore, the expression for  $V(\hat{\lambda}_j^V(s))$  is given by

$$\begin{aligned}
 V(\hat{\lambda}_j^V(s)) &= \hat{\pi}_j(s)^2 V(\hat{\lambda}(s)) + \hat{\lambda}(s)^2 V(\hat{\pi}_j(s)) \\
 &= \hat{\pi}_j(s)^2 \frac{\hat{\lambda}(s)(1-\hat{\lambda}(s))}{n(s)} + \hat{\lambda}(s)^2 \frac{\hat{\pi}_j(s)(1-\hat{\pi}_j(s))}{d(s)} \\
 &= \hat{\pi}_j(s)\hat{\lambda}(s) \left( \frac{\hat{\pi}_j(s)(1-\hat{\lambda}(s))}{n(s)} + \frac{\hat{\lambda}(s)(1-\hat{\pi}_j(s))}{d(s)} \right) \\
 &= \hat{\lambda}_j(s) \left( \frac{d(s)\hat{\pi}_j(s) - d(s)\hat{\pi}_j(s)\hat{\lambda}(s) + n(s)\hat{\lambda}(s) - n(s)\hat{\lambda}(s)\hat{\pi}_j(s)}{n(s)d(s)} \right) \\
 &= \hat{\lambda}_j(s) \left( \frac{d(s)\hat{\pi}_j(s) - d(s)\hat{\lambda}(s) + n(s)\hat{\lambda}(s) - d(s)\hat{\pi}_j(s)}{n(s)d(s)} \right) \\
 &= \hat{\lambda}_j(s) \frac{n(s) - d(s)\hat{\lambda}(s)}{n(s)} = \frac{\hat{\lambda}_j(s)(1-\hat{\lambda}_j(s))}{n(s)} \\
 &= V(\hat{\lambda}_j^C(s))
 \end{aligned}$$

Gaynor et al. (1993) showed in continuous time when competing risks data are analyzed nonparametrically, that the full log-likelihood function is a kernel of a multinomial log-likelihood function as in (2), where the continuous time cause-specific-hazards are approximated with discrete time cause-specific-hazards  $\lambda_j(t)$  at failure times. Therefore, the expression for  $V(\hat{F}_j^C(t))$  that is derived for continuous time competing risks data equally applies in discrete time:

$$V(\hat{F}_j^S(t)) = \sum_{s=1}^t \text{Var}(\hat{S}(s-1)\hat{\lambda}_j(s)) + 2 \sum_{s=1}^{t-1} \sum_{k=s+1}^t \text{Cov}(\hat{S}(s-1)\hat{\lambda}_j(s), \hat{S}(k-1)\hat{\lambda}_j(k))$$

where,

$$\begin{aligned}
 \text{Var}(\hat{S}(s-1)\hat{\lambda}_j(s)) &= \text{Var}(\hat{S}(s-1)\hat{\lambda}_j(s)) \\
 &= (\hat{\lambda}_j(s)\hat{S}(s-1))^2 \left( \frac{n(s) - d(s)}{d(s)n(s)} + \sum_{l=1}^{s-1} \frac{d(l)}{n(l)(n(l) - d(l))} \right) \quad (6)
 \end{aligned}$$

and,

$$\begin{aligned}
 \text{Cov}(\hat{S}(s-1)\hat{\lambda}_j(s), \hat{S}(k-1)\hat{\lambda}_j(k)) &= \text{Cov}(\hat{S}(s-1)\hat{\lambda}_j(s), \hat{S}(k-1)\hat{\lambda}_j(k)) \\
 &= (\hat{\lambda}_j(s)\hat{S}(s-1)\hat{\lambda}_j(k)\hat{S}(k-1)) \\
 &\quad \times \left( -\frac{1}{n(s)} + \sum_{l=1}^{s-1} \frac{d(l)}{n(l)(n(l) - d(l))} \right) \quad (7)
 \end{aligned}$$



To show that  $V(\hat{F}_j^C) = V(\hat{F}_j^V)$ , we need to demonstrate that:

$$V(\hat{S}(s-1)\hat{\lambda}_j(s)) = V(\hat{S}(s-1)\hat{\lambda}(s)\hat{\pi}_j(s))$$

and,

$$\text{Cov}(\hat{S}(s-1)\hat{\lambda}_j(s), \hat{S}(k-1)\hat{\lambda}_j(k)) = \text{Cov}(\hat{S}(s-1)\hat{\lambda}(s)\hat{\pi}_j(s), \hat{S}(k-1)\hat{\lambda}_j(k))$$

Now,

$$V(\hat{S}(s-1)\hat{\lambda}(s)\hat{\pi}_j(s)) = (\hat{S}(s-1)\hat{\lambda}_j(s)\hat{\pi}_j(s))^2 \left( \sum_{l=1}^{s-1} \frac{d_{(l)}}{n_{(l)}(n_{(l)} - d_{(l)})} + \frac{n_{(s)} - d_{(s)}}{d_{(s)}n_{(s)}} + \frac{d_{(s)} - d_{(js)}}{d_{(s)}d_{(js)}} \right) \tag{8}$$

Note that:

$$\frac{(n_{(s)} - d_{(s)})}{d_{(s)}n_{(s)}} + \frac{(d_{(s)} - d_{(js)})}{d_{(js)}d_{(s)}} = \frac{d_{(js)}n_s - d_{(s)}d_{(js)} + n_{(s)}d_{(s)} - n_{(s)}d_{(js)}}{d_{(s)}n_{(s)}d_{(js)}} = \frac{n_{(s)} - d_{(js)}}{n_{(s)}d_{(js)}}$$

Substituting this result in (8) and using the fact that  $\hat{\lambda}_j(s) = \hat{\pi}_j(s)\hat{\lambda}(s)$ , we now have:

$$V(\hat{S}(s-1)\hat{\lambda}(s)\hat{\pi}_j(s)) = (\hat{S}(s-1)\hat{\lambda}_j(s))^2 \left( \sum_{l=1}^{s-1} \frac{d_{(l)}}{n_{(l)}(n_{(l)} - d_{(l)})} + \frac{(n_{(s)} - d_{(js)})}{d_{(js)}n_{(s)}} \right) = V(\hat{S}(s-1)\hat{\lambda}_j(s))$$

We now consider:

$$\begin{aligned} \text{Cov}(\hat{S}(s-1)\hat{\lambda}(s)\hat{\pi}_j(s)\hat{S}(k-1)\hat{\lambda}(k)\hat{\pi}_j(k)) &= (\hat{S}(s-1)\hat{\lambda}(s)\hat{\pi}_j(s)\hat{S}(k-1)\hat{\lambda}(k)\hat{\pi}_j(k)) \\ &\times \left( \sum_{l=1}^{s-1} \frac{d_{(l)}}{n_{(l)}(n_{(l)} - d_{(l)})} - \frac{1}{n_{(s)}} \right) \end{aligned} \tag{9}$$

If we replace  $\hat{\lambda}(\cdot)\hat{\pi}_j(\cdot)$  with  $\hat{\lambda}_j(\cdot)$  in the RHS of (9), then:

$$\text{Cov}(\hat{S}(s-1)\hat{\lambda}_j(s)\hat{S}(k-1)\hat{\lambda}_j(k)) = \text{Cov}(\hat{S}(s-1)\hat{\lambda}_j(s)\hat{S}(k-1)\hat{\lambda}_j(k))$$

This completes the proof that:  $V(\hat{F}_j^C(t)) = V(\hat{F}_j^V(t))$ .

## 6. Conclusion

We have presented the vertical model as a nonparametric model for analysis of discrete time competing risks data. We also demonstrated that the proposed model and the cause-specific-hazards model produce identical estimates. We focussed on the estimates for the

cause-specific-hazards and the cumulative incidence functions. We also showed that the standard errors for the estimates of these quantities were identical under both models. Indeed, it is a roundabout way of estimating the cause-specific-hazards, however, there are cases in practice where these quantities cannot be estimated directly from the data such as when some of the subjects have failed with unknown failure causes. Furthermore, the cause-specific-hazards are not appropriate for application in the presence of a sizable proportion of cured subjects. Nicolaie et al. (2015) have extended the model to handle missing failure causes and the same authors, Nicolaie et al. (2018) have upscaled the model to handle cured subjects. The cause-specific-hazards model cannot handle these data complications. The proposed model, therefore, offers a possibility that the proposed model can also be upscaled to handle these challenges in discrete time. Ndlovu et al. (2020) have presented the vertical model as a nonparametric model for analysis of discrete time data that comes with missing failure causes. Another data complication that has not been explored as yet in the literature is the possibility that data might come with missing failure causes as well as cured subjects.

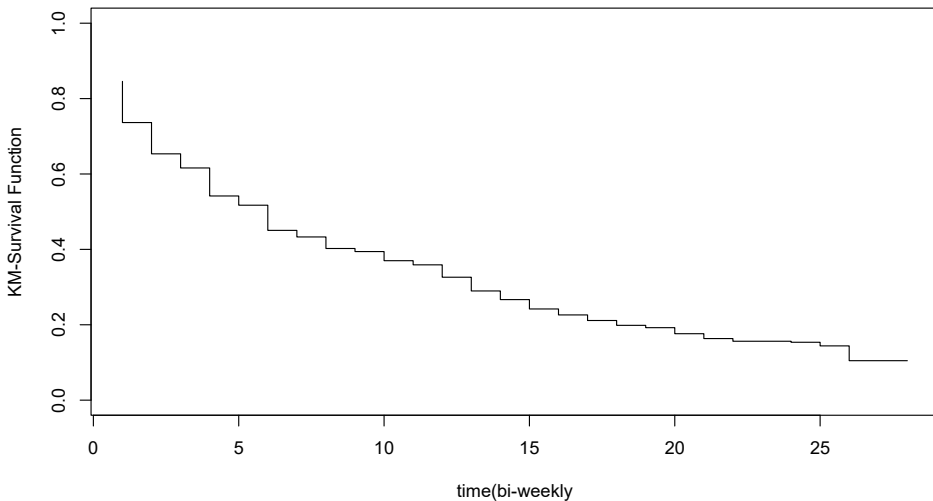


Figure 2: The KM-Survival Function

In clinical trials of a drug for treatment of some cancer, for example, implicit in the study is the expectation that data from that study may have a significant proportion of cured subjects if the drug proves to be effective against the cancer. It is also possible that the failure causes for some of the failures may not be recorded. The subjects with missing failure causes and cured subjects are distinct subjects because cured subjects are assumed to be mixed with censored subjects, whereas missing failure causes relate to subjects that failed. It is, therefore, not inconceivable, that data may come with missing failure causes and cured subjects. In fact, this very data set that was used for illustrative purpose in this article has

missing failure causes and it also presents some evidence that there is a portion of cured subjects, albeit, minimal.

In Figure 2 we have plotted the KM survival function estimate for 3343 subjects. It is evident from the plot that the survival function does not approach zero fast enough, i.e. there is a portion of cured subjects. This means the cause-specific-hazards and cumulative incidence function estimates that were obtained from the proposed model are understated and the extent of bias is directly proportional to the relative size of cured subjects. This is an area that requires further exploration and our opinion is that the vertical model is a strong candidate for handling such data.

## References

- Ambrogi, F., Biganzoli, E., and Boracchi, P., (2009). Estimating Crude Cumulative Incidences through Multinomial Logit Regression on Discrete Cause Specific Hazard. *Computational Statistics and Data Analysis*, 53, pp. 2767–2779.
- Berger, M., Schmid, M., Schmitz-Valckenberg, Welchowski, T., and Bayermann, S., (2020). Subdistribution hazard models for competing risks in discrete time. *Biostatistics*, 21, pp. 449–466.
- Croissant, Y. and Graves, S., (2020). Ecdat: Data Sets for Econometrics. R package version 0.3 – 7
- Davis, T. P. and Lawrance, A. J., (1989). The likelihood for competing risk survival analysis. *Board of the Foundation of the Scandinavian Journal of Statistics*, 16, pp. 23–28.
- Dinse, G. and Larson, M., (1986). A note on semi-Markov models for partially censored data. *Biometrika*, pp. 379–386.
- Gaynor, J. J., Feuer, E. J., Tan, C. C., Wu, D. H., Little, C. R., Straus, D. J., Clarkson, B. D., and Brennan, M. F., (1993). On the use of cause-specific failure and conditional failure probabilities. examples from clinical oncology data. *American Statistical Association*, 88, pp. 400–409.
- Larson, M. G. and Dinse, G. E., (1985). A Mixture Model for the Regression Analysis of Competing Risks Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 34, pp. 201–211.
- Lee, M., Feuer, E., and Fine, J., (2018). On the analysis of discrete time competing risks data. *Biometrics*. doi.org/10.1111/biom.12881.

- Ndlovu, B. D., Melesse, S., and Zewotir, T., (2020). A nonparametric vertical model: An application to discrete time competing risks data with missing failure causes. *South African Journal of Statistics*, 6, pp. 534–545.
- Nicolaie, M., van Houwelingen, H. C., and Putter, H., (2010). Vertical modeling: A pattern mixture approach for competing risks modeling. *Statistics in Medicine*, 29, pp. 1190–1205.
- Nicolaie, M. A., Taylor, J., and Legrand, C., (2018). Vertical modeling: analysis of competing risks data with a cure fraction. *Lifetime Data Analysis*

# Spatial Prediction in Small Area Estimation

Martin Vogt<sup>1</sup>, Partha Lahiri<sup>2</sup>, Ralf Münnich<sup>3</sup>

## ABSTRACT

Small area estimation methods have become a widely used tool to provide accurate estimates for regional indicators such as poverty measures. Recent research has provided evidence that spatial modelling still can improve the precision of regional and local estimates. In this paper, we provide an intrinsic spatial autocorrelation model and prove the propriety of the posterior under a flat prior. Further, we show using the SAIPE poverty data that the gain in efficiency using a spatial model can be essentially important in the presence of a lack of strong auxiliary variables.

**Key words:** Fay-Herriot, CAR, poverty estimation, spatial models.

## 1. Introduction

International programmes such as the United Nations Sustainable Development Goals (SDG), United States Small Area Income and Poverty (SAIPE), the strategy for combating poverty in the European Union need poverty estimates at disaggregated levels for making public policies. Survey data to provide the necessary information on indicators for poverty and social exclusion are generally constructed at regional rather than local levels. Due to budgetary constraints, it is generally not feasible to allocate samples for all conceivable small areas in which different stakeholders may be interested. The estimation for these unplanned small areas may become problematic if the survey does not provide any sample information for these local areas. A standard solution for this problem is to employ a regression method that exploits a possible relationship between the variable of interest and a set of predictor variables available for both planned and unplanned areas. The method essentially generates synthetic estimates that are subject to considerable bias since the method does not use any direct information on the variable of interest for small areas. One potential way to reduce the bias is to utilize data on the variable of interest from neighbouring areas. This can be achieved by incorporating small area specific random effects, which are then linked by a spatial model; see Saei and Chambers (2005). The method is indeed much more complex than the regression method and its success depends on the ability to define a suitable spatial neighbourhood, specification of a spatial model and estimation of additional parameters of the spatial model.

---

<sup>1</sup>Trier University of Applied Sciences, Germany.

E-mail: vogt@hochschule-trier.de. ORCID: <https://orcid.org/0009-0003-2934-1415>.

<sup>2</sup>University of Maryland, College Park, MD 20742, USA.

E-mail: plahiri@umd.edu. ORCID: <https://orcid.org/0000-0002-7103-545X>.

<sup>3</sup>Economics, Economic and Social Statistics Department, Trier University, Germany.

E-mail: muennich@uni-trier.de. ORCID: <https://orcid.org/0000-0001-8285-5667>.

© M. Vogt, P. Lahiri, R. Münnich. Article available under the CC BY-SA 4.0 licence



In the context of mapping the risk from a disease for granular levels, spatial models have been implemented by both empirical Bayes (see, e.g., Clayton and Kaldor, 1987) and hierarchical Bayes (see, Maiti, 1998) methods. Researchers also considered empirical best prediction approach to implement various extensions of the well-celebrated Fay-Herriot small area model that incorporates spatial correlations; see Saei and Chambers , 2005, Singh et al. (2005), Petrucci et al. (2005), Petrucci and Salvati (2006), Petrucci and Salvati (2008), Petrucci and Salvati (2009), and others. For the hierarchical Bayes approach to the spatial Fay-Herriot models, see You and Zhou (2011) and Chung and Datta (2022).

The estimation methodologies developed in the papers cited in the preceding paragraph do not cover the intrinsic CAR model described in Besag et al. (1991) because these papers exclude models with spatial correlation 1. In this paper, we develop a hierarchical Bayes methodology for an extension of the Fay-Herriot model (Fay and Herriot, 1979) that incorporates spatial neighbourhood using a intrinsic CAR model. Thus the proof of posterior propriety for our model does not follow as a corollary of Chung and Datta (2022). Our research is following the PhD thesis of (Vogt , 2010) and is closely linked to the research of Ghosh et al. (1998) and Sun et al. (1999). Note that Sun et al. (1999) extended the research of Ghosh et al. (1998) and show the propriety of the posterior distribution of hierarchical models using CAR(1) distributions. We apply the same spatial structure and the same prior distributions considered by Sun et al. (1999). However, Sun et al. (1999) assume the sampling variances to be *unknown*, but *equal* and, therefore, the model does not include the Fay-Herriot type model with *known* but *unequal* sampling variances. Hence, their theorem does not ensure the propriety of posterior for the model considered in this paper. We adapt the theorem of Sun et al. (1999) to include the spatial general linear mixed model with known but unequal sampling variances.

We apply the proposed model to data of the SAIPE program in the United States. Our application shows that especially if non-sampled areas are present, the incorporation of spatial neighbourhood improves the estimation. The usefulness of the spatial model can also be observed when the quality of auxiliary information is not good, which is often the case in many applications. However, even if the spatial information is already included in the auxiliary variables, the inclusion of the spatial structure does not worsen the results.

The paper is structured as follows. In the next Section, we first obtain a closed-form expression for the Bayes estimator of the small area mean when no sample from the area is available. For this part of the paper, we consider a non-intrinsic CAR model with known hyperparameters. Our analytical calculations allow us to interpret the Bayes estimator and compare it with the alternative synthetic estimator. We then propose an extension of the Fay-Herriot model (see Fay and Herriot, 1979) that incorporates spatial correlations using an intrinsic CAR model. We show the propriety of the posterior for this proposed model under certain regularity conditions. Using the Small Area Income and Poverty (SAIPE) data of the United States Census Bureau (see Bell and Franco , 2017), we demonstrate in the subsequent Section that spatial correlation could considerably improve on the associated hierarchical Bayes methodology if the area specific auxiliary data are either weak or not available. Our investigation reveals that the need for complex spatial models diminishes as strong area specific predictor variables become available.

## 2. Theory

In this section, the Fay-Herriot model is extended by including prior distributions on the regression coefficient and the variance component. Afterward, the independence assumption of the random intercepts is replaced by the conditional autoregressive structure. Then, a formula for the mean of the unsampled area is derived, and finally the propriety of the posterior distribution is proved for our intrinsic CAR Fay-Herriot model.

### 2.1. Spatial Hierarchical Extension of the Fay-Herriot Model

The Fay-Herriot model is given by:

$$\begin{aligned} Y_i | \theta_i &\stackrel{\text{ind}}{\sim} N(\theta_i, \sigma_{\varepsilon,i}^2) \\ \theta_i &= X_i \beta + u_i \\ u_i &\stackrel{\text{ind}}{\sim} N(0, \sigma_u^2 I), \end{aligned} \tag{1}$$

where  $Y_i$  is an estimate of the true small area mean  $\theta_i$ , the sampling variances  $\sigma_{\varepsilon,i}^2$  are assumed to be known.  $X_i$  is  $s \times 1$  vector of known auxiliary variables,  $\beta$  is  $s \times 1$  vector of unknown regression coefficients,  $i = 1, \dots, m$ .

We also consider the following extension of the above model that incorporates possible spatial correlations:

$$\begin{aligned} Y_i | \theta_i &\stackrel{\text{ind}}{\sim} N(\theta_i, \sigma_{\varepsilon,i}^2), \\ \theta_i &= X_i \beta + u_i \\ u &\sim N(0, \sigma_u^2 (I - p\tilde{Q})^{-1} W) \end{aligned} \tag{2}$$

where  $W = \text{diag}(1 / \sum_{j=1}^k Q_{i,j})$  and  $\tilde{Q}_{i,j} = \frac{Q_{i,j}}{\sum_{j=1}^k Q_{i,j}}$  are the weights suggested by Banerjee et al. (2004, p. 79). The neighborhood matrix  $Q$  is symmetric with  $Q_{ii} = 0$ .

The overall goal of this section is to analyze how effective the spatial hierarchical Fay-Herriot model (2) is, in terms of prediction for one unsampled area, compared to the corresponding hierarchical Bayes methodology without spatial correlations. To do this a formula for the mean of the unsampled area is derived in the following section.

### 2.2. The Predicted Mean of One Unsampled Area

In order to derive a formula for the mean of the unsampled area under model (2), the conditional distribution of  $Y_1$  representing the unsampled area given the other areas  $Y_2$  is needed. Assume that the hyperparameters of model (2), i.e.,  $\beta$ ,  $\sigma_u^2$  and  $p$ , are known. Then, model (2) may be written in the form:

$$Y \sim N(\mu, \Sigma) \Leftrightarrow \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N \left( \begin{pmatrix} X_1 \beta \\ X_2 \beta \end{pmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right), \tag{3}$$

where  $\Sigma := \sigma_{\varepsilon,i}^2 I + (I - p\tilde{Q})^{-1} W \sigma_u^2$ . Using standard results, we have:

$$Y_1 \mid (Y_2 = y_2) \sim N(\bar{\mu}, \bar{\Sigma}),$$

$$\text{where } \bar{\mu} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (y_2 - \mu_2)$$

$$\text{and } \bar{\Sigma} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}.$$

The conditional mean  $\bar{\mu}$  may be used to predict one unsampled area. Using the block matrix inversion formula:

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \Rightarrow M^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix} \quad (4)$$

an alternative formulation of  $\bar{\mu}$  can be derived. The new specification shows explicitly how the spatial correlations enter the model.

**Lemma 1** Consider the model (2) with known  $\beta$ ,  $\sigma_{\varepsilon,i}^2$ , and  $\sigma_u^2$  in the form (3). Then, the mean  $\bar{\mu}$  of the conditional distribution of the unsampled area  $Y_1$  given the other areas  $Y_2$ , may be written as:

$$\bar{\mu} = X_1 \beta - \sigma_u^2 B [\sigma_{\varepsilon,22}^2 I_{22} W_{22}^{-1} (D - CA^{-1}B) + \sigma_u^2 I_{22}]^{-1} (y_2 - X_2 \beta), \quad (5)$$

where  $A$  is the 1, 1 element,  $B$  the 1, 2 :  $n$  elements,  $C$  the 2 :  $n$ , 1 and  $D$  the 2 :  $n$ , 2 :  $n$  elements of  $\Sigma_T := I - p\tilde{Q}$ . Thus,  $A$  represents the variance of the first area,  $B$  and  $C$  the correlation between the unsampled and the sampled areas, and  $D$  includes the spatial dependence parameter  $p$  between the sampled areas.

**Proof 1** Model (3) follows with the block matrix inversion formula (4):

$$\begin{aligned} \Sigma &= \sigma_{\varepsilon,i}^2 I + \Sigma_T^{-1} W \sigma_u^2 \\ &= \begin{bmatrix} \sigma_{\varepsilon,11}^2 + ((A - BD^{-1}C)^{-1}) W_{11} \sigma_u^2 & -A^{-1}B(D - CA^{-1}B)^{-1} W_{22} \sigma_u^2 \\ -D^{-1}C(A - BD^{-1}C)^{-1} W_{11} \sigma_u^2 & \sigma_{\varepsilon,22}^2 I_{22} + (D - CA^{-1}B)^{-1} W_{22} \sigma_u^2 \end{bmatrix} \\ &=: \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}. \end{aligned}$$

And thus:

$$\begin{aligned} \bar{\mu} &= X_1 \beta + \Sigma_{12} \Sigma_{22}^{-1} (y_2 - X_2 \beta) \\ &= X_1 \beta - A^{-1}B(D - CA^{-1}B)^{-1} W_{22} \sigma_u^2 [\sigma_{\varepsilon,22}^2 I_{22} + (D - CA^{-1}B)^{-1} W_{22} \sigma_u^2]^{-1} (y_2 - X_2 \beta). \end{aligned}$$

Using the fact that  $M^{-1}N^{-1} = (NM)^{-1}$  with  $M = (D - CA^{-1}B)^{-1} W_{22}$ ,



$N = \left[ \sigma_{\varepsilon,22}^2 I_{22} + (D - CA^{-1}B)^{-1} W_{22} \sigma_u^2 \right]^{-1}$  and  $A^{-1} = 1$  it follows that:

$$\begin{aligned} \bar{\mu} &= X_1 \beta - \sigma_u^2 A^{-1} B \left[ (\sigma_{\varepsilon,22}^2 I_{22} + (D - CA^{-1}B)^{-1} \sigma_u^2 W_{22}) \cdot W_{22}^{-1} (D - CA^{-1}B) \right]^{-1} (y_2 - X_2 \beta) \\ &= X_1 \beta - \sigma_u^2 B \left[ \sigma_{\varepsilon,22}^2 I_{22} W_{22}^{-1} (D - CB) + \sigma_u^2 I_{22} \right]^{-1} (y_2 - X_2 \beta) . \end{aligned}$$

□

The following example clarifies the meaning of formula (5).

**Example 1** In this example the mean  $\bar{\mu}$  of formula (5) will be calculated for a situation with 3 areas, where the first area is unsampled. A nearest neighbor structure is assumed, which means that the first area is a neighbor of the second, the second area is a neighbor of the first and the third area, and finally the third area has got area two as a neighbor (see Figure 1).

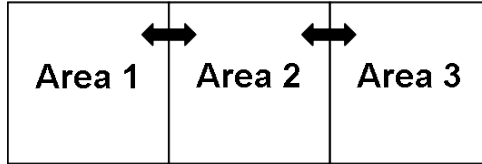


Figure 1: Nearest neighbor structure of 3 areas in a row.

Therefore, the neighborhood matrix  $Q$  is as follows:

$$Q = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} .$$

Dividing each row of  $Q$  by the number of neighbors ( $\tilde{Q}_{i,j} = \frac{Q_{i,j}}{\sum_{j=1}^3 (Q_{i,j})}$ ) yields:

$$\tilde{Q} = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{bmatrix} .$$

The weight matrix  $W = \text{diag}\left(\frac{1}{\sum_{j=1}^3 (Q_{i,j})}\right)$  is given by:

$$W = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} .$$

In the following the mean of one unsampled area shall be calculated using formula (5). Since the first area is unsampled,  $W_{22}$  are the weights for the second and third area, given by

$$W_{22} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix}.$$

Now  $\bar{\mu}$  may be calculated using formula (5). First:

$$\sigma_{\varepsilon,22}^2 I_{22} W_{22}^{-1} = \begin{bmatrix} 2\sigma_{\varepsilon,2}^2 & 0 \\ 0 & \sigma_{\varepsilon,3}^2 \end{bmatrix} \quad (6)$$

and:

$$I - p\tilde{Q} = \begin{bmatrix} 1 & -p & 0 \\ -\frac{p}{2} & 1 & -\frac{p}{2} \\ 0 & -p & 1 \end{bmatrix} =: \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \quad (7)$$

where  $A = 1$ ,  $B = [-p \ 0]$ ,  $C = \left[ \begin{smallmatrix} -\frac{p}{2} \\ 0 \end{smallmatrix} \right]$ , and  $D = \begin{bmatrix} 1 & -\frac{p}{2} \\ -p & 1 \end{bmatrix}$ .

Using formula (7) yields:

$$D - CA^{-1}B = \begin{bmatrix} 1 - p^2/2 & -p/2 \\ -p & 1 \end{bmatrix}. \quad (8)$$

With (6) and (8) it follows that:

$$\sigma_{\varepsilon,22}^2 I_{22} W_{22}^{-1} (D - CA^{-1}B) + \sigma_u^2 I_{22} = \begin{bmatrix} 2\sigma_{\varepsilon,2}^2(1 - p^2/2) + \sigma_u^2 & -\sigma_{\varepsilon,2}^2 p \\ -p\sigma_{\varepsilon,3}^2 & \sigma_{\varepsilon,3}^2 + \sigma_u^2 \end{bmatrix}.$$

Therefore:

$$(\sigma_{\varepsilon,22}^2 I_{22} W_{22}^{-1} (D - CA^{-1}B) + \sigma_u^2 I_{22})^{-1} = \frac{1}{m} \begin{bmatrix} \sigma_{\varepsilon,3}^2 + \sigma_u^2 & \sigma_{\varepsilon,2}^2 p \\ p\sigma_{\varepsilon,3}^2 & 2\sigma_{\varepsilon,2}^2(1 - p^2/2) + \sigma_u^2 \end{bmatrix},$$

where

$$m := (2\sigma_{\varepsilon,2}^2(1 - p^2/2) + \sigma_u^2)(\sigma_{\varepsilon,3}^2 + \sigma_u^2) - \sigma_{\varepsilon,2}^2 \sigma_{\varepsilon,3}^2 p^2.$$

Now all the necessary parts to calculate  $\bar{\mu}$  are derived. Thus:

$$\begin{aligned} \bar{\mu} &= X_1\beta - \sigma_u^2 B [\sigma_{\varepsilon,22}^2 I_{22} W_{22}^{-1} (D - CA^{-1}B) + \sigma_u^2 I_{22}]^{-1} (y_2 - X_2\beta) \\ &= X_1\beta - \sigma_u^2 \begin{bmatrix} -p & 0 \end{bmatrix} \frac{1}{m} \begin{bmatrix} \sigma_{\varepsilon,3}^2 + \sigma_u^2 & \sigma_{\varepsilon,2}^2 p \\ p\sigma_{\varepsilon,3}^2 & 2\sigma_{\varepsilon,2}^2(1 - p^2/2) + \sigma_u^2 \end{bmatrix} (y_2 - \mu_2). \end{aligned}$$

Matrix calculation yields:

$$\bar{\mu} = X_1\beta + \frac{p}{m} \begin{bmatrix} (\sigma_{\varepsilon,3}^2 + \sigma_u^2) & p\sigma_{\varepsilon,2}^2 \end{bmatrix} (y_2 - X_2\beta). \quad (9)$$

Out of formula (9) the following things may be observed.

1. The resulting estimate is a linear combination between the synthetic estimate  $X_1\beta$  and information of the other areas  $(y_2 - X_2\beta)$ .
2. Weight is given to neighbors (area 2) **and** to non-neighbors (area 3).
3. Since  $\sigma_u^2 > 0$  and  $p < 1$  it follows that if  $\sigma_{\epsilon,3}^2$  and  $\sigma_{\epsilon,2}^2$  are of equal size, more power is given to the neighborhood area 2.
4. If  $\sigma_{\epsilon,2}^2$  is large and thus the information of the second area is low, then more strength is taken from area 3 and vice versa.
5. If  $p = 0$  and thus independence is assumed, just the synthetic estimate will be used.

**2.3. Propriety of the Posterior Distribution**

In applications the spatial correlation term  $p$  is frequently assumed to equal 1. Unfortunately, this leads to an improper prior distribution on the random effects  $u$ , the so-called intrinsic CAR model (cf. Besag et al. , 1991 and Besag and Kooperberg , 1995). Thus, the propriety of the posterior distribution is not ensured.

Sun et al. (1999, p. 346) stated the propriety for the intrinsic CAR model with unknown, but equal sampling variances. Let  $Y = (Y_1, \dots, Y_n)$  be the vector of  $n$  observations and let  $X$  and  $Z$  be the  $n \times r$  and  $n \times k$  design matrices. The least squares estimator for  $(\beta', u')$  is given by  $(\hat{\beta}, \hat{u}) = ((X, Z)'(X, Z))^{-1}(X, Z)'Y$ , where  $((X, Z)'(X, Z))^{-1}$  is a generalised inverse of  $(X, Z)'(X, Z)$ . Finally, let  $SSE = Y'\{I_n - (X, Z)((X, Z)'(X, Z))^{-1}\}Y$  be the sum of squared errors. Then, the following theorem holds (cf. Sun et al. , 1999, p. 346).

**Theorem 1** Consider the linear mixed model:

$$Y = X\beta + Zu + \epsilon ,$$

where  $\epsilon \sim N(0, \sigma_\epsilon^2 I)$ . Assume the prior distributions  $f(u) \propto \exp\left(-\frac{1}{2\sigma_u^2}u'Bu\right)$ , where  $B$  is nonnegative definite,  $f(\beta) \propto 1$ ,  $g_\epsilon(\sigma_\epsilon^2) \propto (\sigma_\epsilon^2)^{-(a_\epsilon+1)} \exp(-b_\epsilon/\sigma_\epsilon^2)$  and  $g_u(\sigma_u^2) \propto (\sigma_u^2)^{-(a_u+1)} \exp(-b_u/\sigma_u^2)$ . The variance components are assumed to be a priori independent. Assume the following conditions:

- $rank(X) = r$  and  $rank(u'R_1u + B) = k$ , where  $R_1 = I_n + X(X'X)^{-1}X'$
- $a_u > 0$  and  $b_u > 0$
- $n - r - k - 2a_\epsilon > 0$  and  $SSE + 2b_\epsilon > 0$

Then, the joint posterior distribution of  $(\beta, Z, \sigma_\epsilon^2, \sigma_u^2)$  given  $Y$  is proper.

In this theorem the sampling variances are assumed to be *unknown*, but *equal*. Therefore, this theorem does not ensure the propriety for a Fay-Herriot type model with *known* but *unequal* sampling variances. However, Theorem 1 can be adapted to the spatial general linear mixed model with known, unequal sampling variances, which includes the spatial hierarchical Fay-Herriot model (2).

**Theorem 2** Consider the linear mixed model  $Y = X\beta + Zu + \varepsilon$ , where  $\varepsilon \sim N(0, \Sigma_\varepsilon)$  with known sampling variance matrix  $\Sigma_\varepsilon$ . In addition, the following prior distributions are assumed:  $f(u) \propto \exp\left(-\frac{1}{2\sigma_u^2}u'Bu\right)$ , where  $B$  is nonnegative definite,  $f(\beta) \propto 1$  and  $g(\sigma_u^2) \propto (\sigma_u^2)^{-(a+1)} \exp(-b/\sigma_u^2)$ . Assume the following conditions:

- $\text{rank}(X) = r$  and  $\text{rank}(u'R_1u + B) = k$ , where  $R_1 = \Sigma_\varepsilon^{-1} + \Sigma_\varepsilon^{-1}X(X'\Sigma_\varepsilon^{-1}X)^{-1}X'\Sigma_\varepsilon^{-1}$
- $a > 0$  and  $b > 0$ .

Then, the joint posterior distribution of  $(\beta, u, \sigma_u^2)$  given  $Y$  is proper.

**Proof 2** The idea is to integrate the joint posterior density of  $(\beta, u, \sigma_u^2)$  with respect to the three variables:  $\beta$ ,  $u$ , and  $\sigma_u^2$ . The joint posterior density is proportional to:

$$G := (\sigma_u^2)^{-\frac{1}{2}k} \cdot \exp\left\{-\frac{1}{2}(Y - X\beta - Zu)'\Sigma_\varepsilon^{-1}(Y - X\beta - Zu) - \frac{u^T Bu}{2\sigma_u^2}\right\} \cdot g(\sigma_u^2).$$

The proof is split up into six parts. In parts 1 and 2 the joint posterior is transformed for better handling and integrated with respect to  $\beta$ . Afterward, in parts 2 and 3 the integrated posterior is rearranged to allow for an easier integration with respect to  $u$ . Finally, in parts 5 and 6 the joint posterior is bounded and thus the existence is shown.

1. First,  $G$  is transformed since this helps to better handle the integration with respect to  $\beta$ . This is done by adding and subtracting  $X\hat{\beta}$  and  $Z\hat{u}$ . It follows that:

$$\begin{aligned} & (Y - X\beta - Zu)'\Sigma_\varepsilon^{-1}(Y - X\beta - Zu) \\ &= (Y - X\hat{\beta} - Z\hat{u} - X(\beta - \hat{\beta}) - Z(u - \hat{u}))'\Sigma_\varepsilon^{-1} \cdot \\ & \quad \cdot (Y - X\hat{\beta} - Z\hat{u} - X(\beta - \hat{\beta}) - Z(u - \hat{u})). \end{aligned} \quad (10)$$

Expanding (10) yields:

$$\begin{aligned} & (Y - X\beta - Zu)'\Sigma_\varepsilon^{-1}(Y - X\beta - Zu) \\ &= e'\Sigma_\varepsilon^{-1}e \\ & \quad - e'\Sigma_\varepsilon^{-1}X(\beta - \hat{\beta}) + \\ & \quad + (\beta - \hat{\beta})'X'\Sigma_\varepsilon^{-1}X(\beta - \hat{\beta}) - \\ & \quad - (\beta - \hat{\beta})'X'\Sigma_\varepsilon^{-1}e + \\ & \quad + (\beta - \hat{\beta})'X'\Sigma_\varepsilon^{-1}Z(u - \hat{u}) + (u - \hat{u})'Z'\Sigma_\varepsilon^{-1}X(\beta - \hat{\beta}) - \\ & \quad - e'\Sigma_\varepsilon^{-1}Z(u - \hat{u}) - \\ & \quad - (u - \hat{u})'Z'\Sigma_\varepsilon^{-1}e + \\ & \quad + (u - \hat{u})'Z'\Sigma_\varepsilon^{-1}Z(u - \hat{u}), \end{aligned} \quad (11)$$

where  $e := Y - X\hat{\beta} - Z\hat{u}$ .

2. Now all of the factors in (11) containing  $(\beta - \hat{\beta})$  are collected:

$$(Y - X\beta - Zu)'\Sigma_\varepsilon^{-1}(Y - X\beta - Zu) = (\beta - \hat{\beta} - C_0 - C_1)'X'\Sigma_\varepsilon^{-1}X(\beta - \hat{\beta} - C_0 - C_1) + \text{Const}_0,$$

where

$$C_0 = (X' \Sigma_\varepsilon^{-1} X)^{-1} X' \Sigma_\varepsilon^{-1} e,$$

$$C_1 = (X' \Sigma_\varepsilon^{-1} X)^{-1} X' \Sigma_\varepsilon^{-1} Z(u - \hat{u}).$$

Note that  $X' \Sigma_\varepsilon^{-1} X$  is symmetric.  $Const_0$  is a constant which contains all the factors independent of  $\beta$  :

$$\begin{aligned} Const_0 &= (u - \hat{u})' Z' \Sigma_\varepsilon^{-1} Z(u - \hat{u}) - (u - \hat{u})' Z' \Sigma_\varepsilon^{-1} e - \\ &- e' \Sigma_\varepsilon^{-1} Z(u - \hat{u}) - C_1' X' \Sigma_\varepsilon^{-1} X C_0 - \\ &- C_1' X' \Sigma_\varepsilon^{-1} X C_1 - C_0' X' \Sigma_\varepsilon^{-1} X C_1 - C_0' X' \Sigma_\varepsilon^{-1} X C_0 \end{aligned}$$

Integrating  $G$  with respect to  $\beta$  yields:

$$\int_{\mathbb{R}^p} G \, d\beta = \frac{(2\pi)^{\frac{1}{2}p} |X' \Sigma_\varepsilon^{-1} X|^{-\frac{1}{2}}}{(\sigma_u^2)^{\frac{1}{2}k}} \exp \left\{ -\frac{1}{2} Const_0 - \frac{u' B u}{2\sigma_u^2} \right\} \cdot g(\sigma_u^2). \quad (12)$$

3. Now, the integration with respect to  $u$  is prepared. Therefore, the exponential function (12) is calculated and transformed by collecting the terms containing  $u$  in  $Const_0$  :

- (a)  $C_0' X' \Sigma_\varepsilon^{-1} X C_0$  which is independent of all integration variables and will be seen as a constant.
- (b)  $C_0' X' \Sigma_\varepsilon^{-1} X C_1 = ((X' \Sigma_\varepsilon^{-1} X)^{-1} X' \Sigma_\varepsilon^{-1} e)' X' \Sigma_\varepsilon^{-1} Z(u - \hat{u})$
- (c)  $C_1' X' \Sigma_\varepsilon^{-1} X C_1 = (u - \hat{u})' Z' \Sigma_\varepsilon^{-1} X (X' \Sigma_\varepsilon^{-1} X)^{-1} X' \Sigma_\varepsilon^{-1} Z(u - \hat{u})$
- (d)  $C_1' X' \Sigma_\varepsilon^{-1} X C_0 = -((X' \Sigma_\varepsilon^{-1} X)^{-1} X' \Sigma_\varepsilon^{-1} Z(u - \hat{u}))' X' \Sigma_\varepsilon^{-1} e$
- (e)  $e' \Sigma_\varepsilon^{-1} Z(u - \hat{u})$
- (f)  $(u - \hat{u})' Z' \Sigma_\varepsilon^{-1} e$
- (g)  $(u - \hat{u})' Z' \Sigma_\varepsilon^{-1} Z(u - \hat{u})$

This leads to:

$$Const_0 = (u - \hat{u} - C_2)' Z' R_1 Z(u - \hat{u} - C_2) + Const_1, \quad (13)$$

where

$$C_2 = (Z' R_1 Z)^{-1} Z' (R_1 + \Sigma_\varepsilon^{-1}) e$$

includes the above terms with one  $(u - \hat{u})$ . The constant  $Const_1$  contains all of the terms independent of  $u$  which arise by including  $C_2$  in the formula.

4. Now,  $\frac{u' B u}{\sigma_u^2}$  is considered. Therefore, formula (13) of the previous step is adapted:

$$Const_0 + \frac{u' B u}{\sigma_u^2} = (u - \hat{u} - C_2)' Z' R_1 Z(u - \hat{u} - C_2) + Const_1 + \frac{u' B u}{\sigma_u^2}.$$

This leads to

$$u'Z'R_1Zu + \frac{u'Bu}{\sigma_u^2} + \text{other terms.} \quad (14)$$

A rearrangement of terms in (14) yields:

$$(u - C_3)'R_2(u - C_3) - (C_3)'R_2C_3, \quad (15)$$

where  $R_2 = Z'R_1Z + \frac{B}{\sigma_u^2}$  and  $C_3 = R_2^{-1}Z'R_1Z(\hat{u} + C_2)$ .

Using the integrated  $G$  in (12) together with (15), it follows that:

$$\int_{\mathbb{R}^k} \int_{\mathbb{R}^r} G \, d\beta du = \frac{(2\pi)^{\frac{1}{2}(r+k)} |X'\Sigma_\varepsilon^{-1}X|^{-\frac{1}{2}}}{(\sigma_u^2)^{\frac{1}{2}k+a+1} |R_2|^{\frac{1}{2}}} \exp \left\{ -(C_3)'R_2C_3 - \frac{b}{\sigma_u^2} \right\}. \quad (16)$$

Integration leads to the factor  $|R_2|^{-\frac{1}{2}}$  in front of the exponential function. Since  $Z'R_1ZR_2^{-1}Z'R_1Z$  in  $(C_3)'R_2C_3$  is nonnegative definite this term can be bounded by 0 and thus discarded out of the integral.

5. Using arguments similar to Sun et al. (1999, p. 346), we get:

$$\begin{aligned} |R_2|^{-\frac{1}{2}} &\leq \{ \min(1, (\sigma_u^2)^{-1})^k \cdot |Z'R_1Z + B| \}^{-\frac{1}{2}} \\ &< (1 + (\sigma_u^2)^{\frac{k}{2}}) \cdot |Z'R_1Z + B|^{-\frac{1}{2}}. \end{aligned} \quad (17)$$

6. Finally, combining (16) and (17), we have:

$$\int_{\mathbb{R}^k} \int_{\mathbb{R}^r} G \, d\beta du \leq (2\pi)^{r+k} |X'\Sigma_\varepsilon^{-1}X|^{-\frac{1}{2}} |Z'R_1Z + B|^{-\frac{1}{2}} (J_1 + J_2),$$

where

$$J_1 = \frac{1}{(\sigma_u^2)^{\frac{1}{2}k+a+1}} \exp \left( -\frac{b}{\sigma_u^2} \right),$$

and

$$J_2 = \frac{1}{(\sigma_u^2)^{a+1}} \exp \left( -\frac{b}{\sigma_u^2} \right).$$

Since  $a > 0$  and  $b > 0$  the integrals  $J_1$  and  $J_2$  exist with respect to  $\sigma_u^2$ .

This completes the proof. □

### Remark:

1. In the theorem presented in Sun et al. (1999), the sum of squared errors SSE arises in the assumptions. This term is not important in our case, since it depends only on the known variance-covariance matrix  $\Sigma_\varepsilon$ .
2. In the original proof, all terms containing  $e$  cancel out of formula (11), since SSE is orthogonal to  $e$ . This needs not be the case if  $\Sigma_\varepsilon$  is included in the term.

### 3. An empirical comparison of the intrinsic CAR Fay-Herriot model and non-spatial Fay-Herriot models

In the following, we compare the Fay-Herriot model (1) with the intrinsic CAR Fay-Herriot model (2) (i.e. with  $p = 1$ ) in terms of prediction using real data. To implement the intrinsic CAR Fay-Herriot model, we used priors described in Theorem 2. As for the implementation of the Fay-Herriot model (1), we considered the following prior distributions proposed by Sun et al. (1999):

$$\begin{aligned} \beta &\propto \text{Uniform}(\mathbb{R}^s), \mathbb{R}^s \text{ being the } s\text{-dimensional Euclidean space,} \\ 1/\sigma_u^2 &\sim \Gamma(0.5, 0.005). \end{aligned}$$

#### 3.1. Data

We consider the data from the Small Area Income and Poverty Estimates of the U.S. Census Bureau for the years 1989 and 1993. For details on the data, see Bell and Franco (2017). Four covariates are available: Internal Revenue Service (IRS) pseudo child poverty rate ( $x_1$ ), IRS non-filer rate ( $x_2$ ), food stamp participation rate ( $x_3$ ) and census residuals ( $x_4$ ). The known sampling variances are denoted by  $d$ . In this application the official Small Area Income and Poverty Estimates (SAIPE) estimates are treated as a gold standard. In the application 48 contiguous United States and the District of Columbia are considered. Every area is left out once and is predicted by means of spatial and non-spatial models. This procedure is repeated for different numbers of covariates (0–4) and for the years 1989 and 1993. The estimation results are compared to the official estimates (treated as gold standard in this paper).

#### 3.2. Results

Since the results are similar for the years 1989 and 1993 the following interpretation will just be for the year 1993. Table 1 contains the simulation results for the year 1993 and Table 2 for the year 1989. Column 1 of Table 1 shows different measures of comparison, based on the squared deviance between the estimator and the official value, the absolute deviance and the maximum of the deviance. Since each of the 49 states is left out once, the deviances are averaged over all states. The deviances are constructed for the model containing all of the four covariates, no covariate or each of the covariates alone. Columns 2 and 3 contain the corresponding values for the spatial and non-spatial models. The last two columns compare the deviances of the spatial and non-spatial model with each other (difference and ratio). The following observations can be made:

1. If no covariates are included, the deviances for the spatial and non-spatial models are large. These values decrease as the quality and number of covariates increases. The lowest value is reached, when all covariates are included.
2. The ratio of the non-spatial deviance compared to the spatial is large, if no or weak covariates are included. The ratio decreases, if the quality of the covariates improves.

Table 1: Simulation results for the year 1993 .

	<b>spatial</b>	<b>non-sp.</b>	<b>spatial – non-sp.</b>	<b><math>\frac{\text{non-sp.}}{\text{spatial}}</math></b>
$\frac{1}{49} \cdot \sum_{i=1}^{49} ((\text{Estimator}_i - \text{Official}_i)^2)$				
all	1,08	1,08	0,00	1,00
x1	7,76	12,79	-5,03	1,65
x2	17,04	27,44	-10,40	1,61
x3	4,74	4,63	0,11	0,98
x4	23,42	33,55	-10,13	1,43
without	25,14	34,01	-8,86	1,35
$\frac{1}{49} \cdot \sum_{i=1}^{49} ( \text{Estimator}_i - \text{Official}_i )$				
all	0,79	0,79	0,00	1,00
x1	2,20	2,84	-0,65	1,29
x2	3,00	4,04	-1,05	1,35
x3	1,76	1,78	-0,02	1,01
x4	3,76	4,85	-1,09	1,29
without	3,85	4,89	-1,04	1,27
<b>max( Estimator – Official )</b>				
all	0,07	0,07	0,00	1,00
x1	0,12	0,17	-0,05	1,38
x2	0,24	0,30	-0,05	1,21
x3	0,12	0,12	0,00	0,99
x4	0,26	0,29	-0,03	1,12
without	0,29	0,27	0,02	0,93

3. If all covariates are included there is no gain by using the spatial model.

The same effects can be seen, in Figures 2, 4 and 5. These figures compare the predicted values of the spatial and non-spatial models with varying numbers of covariates. Figure 2 shows that if all covariates are included, there is no visible difference between the spatial and non-spatial models. However, if no covariate is included, then the predicted values of the non-spatial model compared to the official values are almost constant. But the spatial model improves the relationship. The same effect can be observed if covariates of different quality are included (Figures 4 and 5). Figure 3 underlines these results, by showing the squared deviance of the spatial and non-spatial models for all 4 covariates (upper plots) and no covariates (lower plots) on the map. If no covariates are included the spatial model performs better than the non-spatial model. This effect diminishes if all covariates are included.



Table 2: Simulation results for the year 1989 .

	<b>spatial</b>	<b>non-sp.</b>	<b>spatial – non-sp.</b>	<b><math>\frac{\text{non-sp.}}{\text{spatial}}</math></b>
$\frac{1}{49} \cdot \sum_{i=1}^{49} ((\text{Estimator}_i - \text{True}_i)^2)$				
all	0,97	0,95	0,02	0,98
x1	4,53	4,81	-0,28	1,06
x2	12,14	21,83	-9,69	1,80
x3	3,89	5,14	-1,25	1,32
x4	17,27	27,38	-10,11	1,59
without	16,11	26,57	-10,46	1,65
$\frac{1}{49} \cdot \sum_{i=1}^{49} ( \text{Estimator}_i - \text{True}_i )$				
all	0,82	0,81	0,01	0,98
x1	1,51	1,55	-0,04	1,02
x2	2,64	3,42	-0,79	1,30
x3	1,58	1,81	-0,23	1,15
x4	3,13	4,04	-0,91	1,29
without	3,16	3,97	-0,81	1,26
<b>max( Estimator – True )</b>				
all	0,05	0,05	0,00	1,00
x1	0,14	0,14	0,00	1,02
x2	0,25	0,33	-0,08	1,30
x3	0,09	0,10	-0,01	1,07
x4	0,30	0,34	-0,04	1,12
without	0,25	0,34	-0,09	1,34

#### 4. Conclusion

In this paper, we have developed a hierarchical Bayes methodology for an extension of the well-celebrated Fay-Herriot model that incorporates spatial correlation using an intrinsic CAR model. We have proved the propriety of the posterior distribution for our proposed model. We have tested the effect of covariates on the estimation results. An application to SAIPE data revealed that modeling spatial correlation can considerably improve on the associated hierarchical Bayes methodology if the area specific auxiliary data are either weak or not available. This effect diminishes if the quality of the area specific covariates improves. Like Besag et al. (1991) we also assumed proper prior for the variance component. As for the future research, it will be of interest to study the sensitivity of such proper prior and to compare our hierarchical Bayes estimator with various empirical best predictors and hierarchical Bayes estimators proposed in the literature that use non-intrinsic CAR model extensions of the Fay-Herriot model.

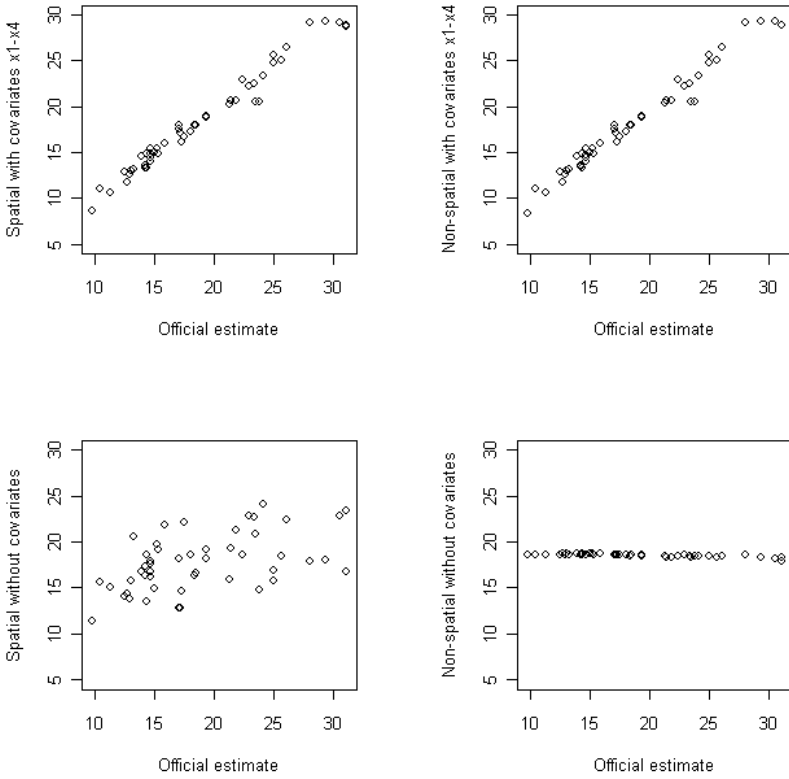


Figure 2: Predicted values of the spatial and non-spatial FH model compared to the official estimates 1993: 4 and no covariates

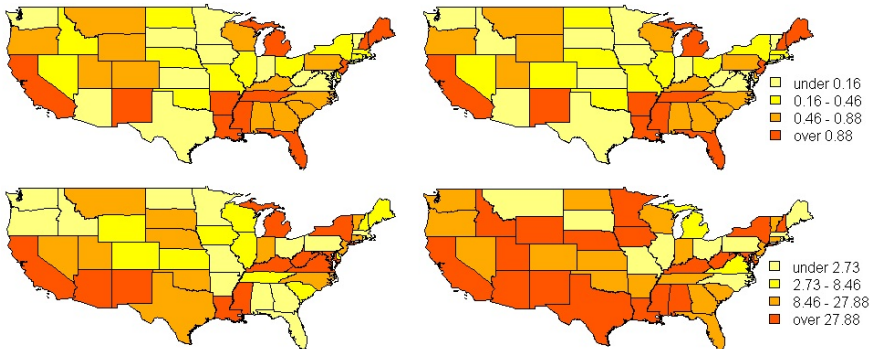


Figure 3: Squared deviance of the spatial and non-spatial FH model for 4 (upper plots) and no (lower plots) covariates.

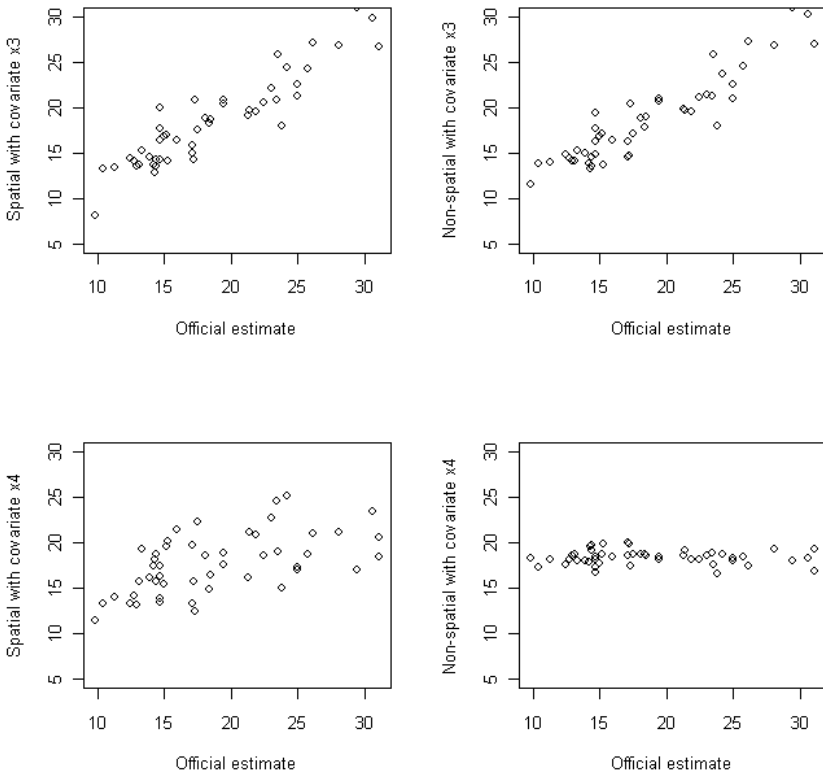


Figure 4: Predicted values of the spatial and non-spatial FH model compared to the official estimates 1993: covariates x3, x4

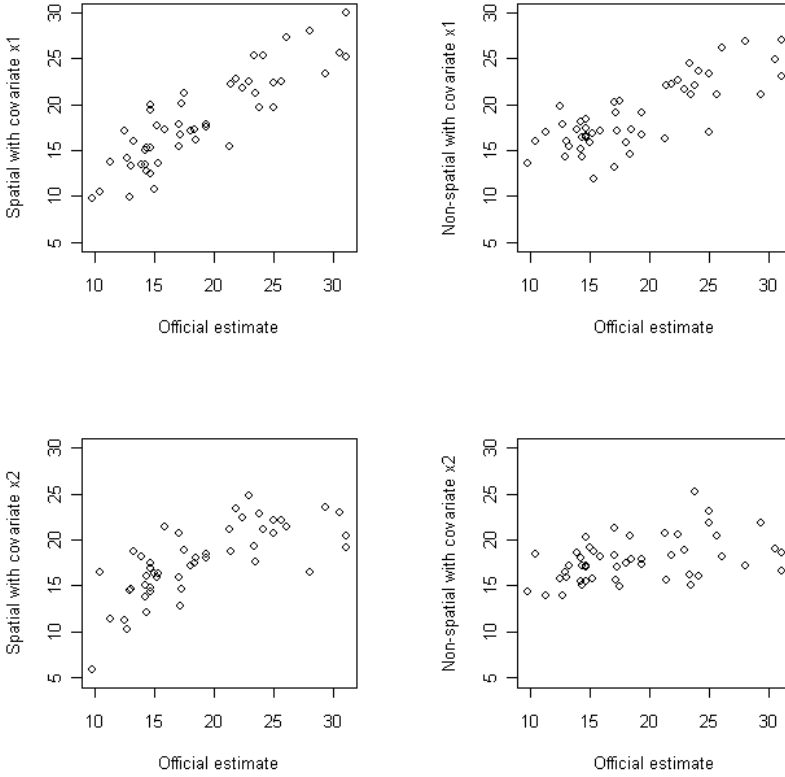


Figure 5: Predicted values of the spatial and non-spatial FH model compared to the official estimates 1993: covariates x1, x2

## References

- Banerjee, S., Carlin, B. and Gelfand, A. E., (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: Chapman & Hall/CRC.
- Bell, W. R., and Franco, C., (2017). Small Area Estimation – State Poverty Rate Model Research Data Files. Available at [https:// www.census.gov/srd/csrreports/byyear.html](https://www.census.gov/srd/csrreports/byyear.html) [accessed October 22, 2018]
- Besag, J., York, J. and Mollie, A., (1991). Bayesian image restoration with two applications in spatial statistics, *Annals of the Institute of Statistical Mathematics*, 43, 1–59 (With discussion).
- Besag, J. and Kooperberg, C. L., (1995). On conditional and intrinsic autoregressions, *Biometrika*, 82, 733–746.
- Chung, H. C. and Datta, G. S., (2022). Bayesian spatial models for estimating means of sampled and non-sampled small areas, *Survey Methodology*, 48, 2, 463–489.
- Clayton, D. and Kaldor, J., (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics*, 43, 671–681.
- Cressie, N. A. C., (1993). *Statistics for spatial data*, New York: John Wiley & Sons.
- Fay, R. E. and Herriot, R. A., (1979). Estimates of Income for Small Places: An Application of James Stein Procedures to Census Data, *Journal of the American Statistical Association*, 74 (366), 269–277.
- Geobugs user manual version 1.2, (2004).
- Ghosh, M., Natarajan, K., Stroud, T. W. F. and Carlin B. P., (1998). Generalized Linear Models for Small-Area Estimation, *Journal of the American Statistical Association*, 93, 273–282.
- Kass, R. E. and Wassermann, L., (1996). The selection of prior distributions by formal rules, *Journal of the American Statistical Association*, 91, 1343–1370.
- Kelsall, J. E. and Wakefield, J. C., (1999). Discussion of Bayesian models for spatially correlated disease and exposure data. by Best et al. in *Bayesian Statistics 6*. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds), Oxford: Oxford University Press
- Jiang, J. and Lahiri, P., (2006). Mixed Model Prediction and Small Area Estimation, *Test*, 15 (1), 1–96.
- Maiti, T., (1998). Hierarchical Bayes estimation of mortality rates for disease mapping, *Journal of Statistical Planning and Inference*, 69, 339–348.
- Petrucci, A. Pratesi, M. and Salvati, N., (2005). Geographic Information in Small Area Estimation: Small Area Models and Spatially Correlated Random Area Effects, *Statistics in Transition*, 3.

- Petrucci, A. and Salvati, N., (2006). Small Area Estimation for Spatial Correlation in Watershed Erosion Assessment, *Journal of Agricultural, Biological, and Environmental Statistics*, 11, 2, 169–182.
- Petrucci, A. and Salvati, N., (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects, *Statistical Methods and Applications*, 17, 113–141.
- Petrucci, A. and Salvati, N., (2009). Small Area Estimation in the Presence of Correlated Random Area Effects, *Journal of Official Statistics*, 25, 1, 37–53.
- Rao, J. N. K., (2003). *Small Area Estimation*, New York: John Wiley & Sons.
- Saei, A. and Chambers, R., (2005). Out of Sample Estimation for Small Areas using Area Level Data, Southampton Statistical Sciences Research Institute Methodology Working Paper M05/11.
- Singh, B. B., Shukla, G. K. and Kundu, D., (2005). Spatio-Temporal Models in Small Area Estimation, *Survey Methodology*, 31, 2, 183–195.
- Sun, D., Tsutakawa, R. K. and Speckman, P. L., (1999). Posterior Distribution of hierarchical models using CAR(1) distributions, *Biometrika*, 86, 341–350.
- You, Y. and Zhou, Q. M., (2011). Hierarchical Bayes small area estimation under a spatial model with application to health survey data, *Survey Methodology*, 37, 1, 25–37.
- Vogt, M., (2010). *Bayesian Spatial Modeling: Propriety and Applications to Small Area Estimation with Focus on the German Census 2011*. PhD thesis, Universität Trier. DOI: 10.25353/ubtr-xxxx-2ba6-6f2e/.

# Efficient estimation of population mean in the presence of non-response and measurement error

Kuldeep Kumar Tiwari<sup>1</sup>, Vishwantra Sharma<sup>2</sup>

## ABSTRACT

In real-world surveys, non-response and measurement errors are common, therefore studying them together seems rational. Some population mean estimators are modified and studied in the presence of non-response and measurement errors. Bias and mean squared error expressions are derived under different cases. For all estimators, a theoretical comparison is made with the sample mean per unit estimator. The Monte-Carlo simulation is used to present a detailed picture of all estimators' performance.

**Key words:** non-response, measurement error, mean squared error, efficiency, mean estimation.

## 1. Introduction

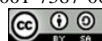
The sampling technique is the most effective way to make population predictions by using a sample of units from the population. All of the survey's sampling strategies are theoretically based on some assumptions. One assumption that almost never holds true in real-world surveys is that all units in the sample will respond. Non-response could be due to a variety of factors, including the respondent's availability, discomfort with the questions/interviewer, or a lack of desire to contribute. However, the increase in error caused by non-response had a significant impact on the final results. In the presence of non-response in sample surveys, Hansen and Hurwitz (1946) proposed a method for estimating the population mean. They usually use it for mail surveys because they are less expensive. To use this method, first send a questionnaire to all of the units in the sample via mail. After that, select a sub-sample from the non-respondent units and conduct a direct or telephone interview with them. When contacted directly, he assumes that every unit in the non-respondent sub-sample responds. Hansen and Hurwitz (1946) defined the estimator of population mean in the presence of non-response as  $\bar{y}_t^* = \left(\frac{n_1}{n}\right)\bar{y}_{n_1} + \left(\frac{n_2}{n}\right)\bar{y}_r$ ; where  $\bar{y}_{n_1} = \frac{1}{n_1} \sum_{i=1}^{n_1} y_i$ ,  $\bar{y}_r = \frac{1}{r} \sum_{i=1}^r y_i$ ,  $n$  is sample size,  $n_1$  is the number of respondent in the sample and  $n_2$  is the number of non-respondent such that  $n_1 + n_2 = n$ .  $r = \frac{n_2}{k}$ ,  $k > 0$  is the size of sub-sample of non-respondent. Using this concept of handling non-response, many authors have proposed estimators for a variety of cases over decades. Some of them are Cochran (1977), Rao (1986), Okafor and Lee (2000), Kreuter et al. (2010), Khan et al. (2014), Luengo (2016), Khare and Sinha (2019), Sharma and Kumar (2020), Pandey et al. (2021), Sinha et al. (2022).

<sup>1</sup>Department of Mathematics, Chandigarh University, Mohali, Punjab, India.

E-mail: kuldeep.smvd@gmail.com. ORCID: <https://orcid.org/0000-0001-5083-1206>.

<sup>2</sup>Directorate of Census Operations, Jammu and Kashmir, India.

E-mail: vishwantrasharma07@gmail.com. ORCID: <https://orcid.org/0000-0001-7387-0670>.



Aside from non-response, measurement error is another error that affects the results of a real-life survey. We assume that all of the data that have been recorded and processed are accurate. However, in real-life surveys, this is purely hypothetical. Measurement error can be caused by a variety of factors, including interviewer bias, respondent bias, error in recording and processing the data, and so on. Some notable works on the estimation in the presence of measurement error are Cochran (1977), Fuller (1987), Shalabh (1997), Srivastava and Shalabh (2001), Gregoire and Salas (2009), Diane and Giordan (2012), Tiwari et al. 2022.

Since the presence of non-response and measurement error is expected to be in any survey, so it is desirable to study both of them at the same time. Very few works have been done so far on this. The contribution of researchers in this area is Kumar et al. (2015), Singh and Sharma (2015), Azeem and Hanif (2016), Kumar and Bhoulal (2018), Kumar et al. (2018), Singh et al. (2018), Zahid et al. (2022), Tiwari et al. (2022).

So, here we carried out a study on the estimation of population mean in the presence of non-response and measurement error.

## 2. Notations

Let a finite population of size  $N$  be divided into two groups as respondent of size  $N_1$  and non-respondent of size  $N_2$ . Let a sample of size  $n$  be taken from the population among which  $n_1$  are respondent and  $n_2$  are non-respondent. A sub-sample of size  $r (= \frac{n_2}{k})$ ,  $k > 0$  is taken from  $n_2$  non-respondents. At  $i^{th}$  unit of population,  $y_i$  and  $x_i$  be the observed values of study and auxiliary variables and  $y_{ti}$ ,  $x_{ti}$  be their true values, respectively.

The other notations are  $\bar{x}_i^* = (\frac{n_1}{n})\bar{x}_{n_1} + (\frac{n_2}{n})\bar{y}_r$ ,  $\bar{x}_{n_1} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i$ ,  $\bar{x}_r = \frac{1}{r} \sum_{i=1}^r x_i$ ,  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ ,  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ .

The population mean and variance for  $y$  are  $\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$ ,  $S_Y^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2$ . The population mean and variance for  $x$  are  $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$ ,  $S_X^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$ . Population variance for  $y$  and  $x$  for the group of non-respondent is

$S_{Y(2)}^2 = \frac{1}{N_2-1} \sum_{i=1}^{N_2} (Y_i - \bar{Y})^2$  and  $S_{X(2)}^2 = \frac{1}{N_2-1} \sum_{i=1}^{N_2} (X_i - \bar{X})^2$  respectively.

Let  $U_i = y_i - y_{ti}$  and  $V_i = x_i - x_{ti}$  be the measurement error on the study and auxiliary variable respectively at  $i^{th}$  unit of the population. So,  $\bar{y}^* = \bar{y}_i^* + \bar{U}^*$  and  $\bar{x}^* = \bar{x}_i^* + \bar{V}^*$ , where  $\bar{U}^* = (\frac{n_1}{n})\bar{U}_{n_1} + (\frac{n_2}{n})\bar{U}_r$ ,  $\bar{U}_{n_1} = \frac{1}{n_1} \sum_{i=1}^{n_1} U_i$ ,  $\bar{U}_r = \frac{1}{r} \sum_{i=1}^r U_i$  and  $\bar{V}^* = (\frac{n_1}{n})\bar{V}_{n_1} + (\frac{n_2}{n})\bar{V}_r$ ,  $\bar{V}_{n_1} = \frac{1}{n_1} \sum_{i=1}^{n_1} V_i$ ,  $\bar{V}_r = \frac{1}{r} \sum_{i=1}^r V_i$ .

Study variable  $y$  and auxiliary variable  $x$  are correlated with correlation coefficient  $\rho$  and  $\rho_2$  is the correlation coefficient between  $y$  and  $x$  for the group of non-respondent. Since there is no relationship between measurement errors occurring on  $y$  and  $x$ , so  $U$  and  $V$  must be independent. Also, since there will be both under-reporting and over-reporting in measurement error, so we assume that mean of  $U$  and mean of  $V$  are zero. The population variance of measurement error associated with  $y$  is  $S_U^2 = \frac{1}{N-1} \sum_{i=1}^N (U_i - \bar{U})^2$  and the population variance of measurement error associated with  $x$  is  $S_V^2 = \frac{1}{N-1} \sum_{i=1}^N (V_i - \bar{V})^2$ . Population variances of  $U$  and  $V$  for the group of non-respondent are  $S_{U(2)}^2 = \frac{1}{N_2-1} \sum_{i=1}^{N_2} (U_i - \bar{U})^2$  and  $S_{V(2)}^2 = \frac{1}{N_2-1} \sum_{i=1}^{N_2} (V_i - \bar{V})^2$  respectively.  $b_{yx}$  is sample regression coefficient of  $y$  on  $x$ .



Other notations used in the article are:  $\theta = \frac{W_2(k-1)}{n}$ ,  $W_2 = \frac{N_2}{N}$ ,  $\lambda = \frac{1}{n} - \frac{1}{N}$ ,  $R = \frac{\bar{Y}}{\bar{X}}$ ,  $w_1 = \frac{n_1}{n}$ ,  $w_2 = \frac{n_2}{n}$ .

### 3. Brief review of literature

In this section, we discuss some estimators from the literature that will be further used.

Searls (1964) proposes an estimator  $t_1 = k\bar{y}$  in simple random sampling, where  $k$  is a suitable constant. He shows that  $MSE(t_1)$ , that is mean square error of  $t_1$  is less than the variance of  $\bar{y}$ , hence  $t_1$  is preferable over usual estimator  $\bar{y}$ . Cochran (1940) defined the ratio estimator  $t_2 = \bar{y}(\frac{\bar{X}}{\bar{x}})$ . He further studied the ratio estimator in the presence of non-response when non-response occurs on both study variables and auxiliary variables in Cochran (1977). Rao (1986) studied the ratio estimator when there is non-response only on the study variable. Shalabh (1997) adapted the ratio estimator and presented a study on the ratio method of estimation in the presence of measurement error. Murthy (1964) proposes the product method of estimation by defining the estimator  $t_3 = \bar{y}(\frac{\bar{x}}{\bar{X}})$ . He shows that the product method of estimation is better to use when there is a high negative correlation between the study and the auxiliary variable. Khare and Srivastava (1993) studied ratio and product estimator in double sampling when there is non-response on both study and auxiliary variables. Cochran (1977) studied the usual regression estimator  $t_4 = \bar{y} + b_{yx}(\bar{X} - \bar{x})$  and its properties when there is non-response on both study and auxiliary variables. Srivastava and Shalabh (2001) examined the regression estimator in the presence of measurement error. Okafor and Lee (2000) presented a study on ratio and regression estimator when there is non-response on both variables in the double sampling scheme. Srivastava (1967) generalise the ratio estimator by proposing  $t_5 = \bar{y}(\frac{\bar{X}}{\bar{x}})^\alpha$ . Rao (1991) proposed a difference estimator  $t_6 = k_1\bar{y} + k_2(\bar{X} - \bar{x})$  in simple random sampling and shows that it works better than regression estimator. Bahl and Tuteja (1991) first time uses exponential function to estimate the population mean by defining ratio and product type estimator  $t_7 = \bar{y}\exp(\frac{\bar{X}-\bar{x}}{\bar{X}+\bar{x}})$  and  $t_8 = \bar{y}\exp(\frac{\bar{x}-\bar{X}}{\bar{x}+\bar{X}})$  respectively. Using ratio and product estimator, Singh and Espejo (2003) proposed an estimator as  $t_9 = \bar{y}[a(\frac{\bar{X}}{\bar{x}}) + (1-a)(\frac{\bar{x}}{\bar{X}})]$  and show that its optimum mean square error (MSE) is the same as regression estimator. Kadilar and Cingi (2004) proposed an estimator using regression and ratio estimator as  $t_{10} = [\bar{y} + b_{yx}(\bar{X} - \bar{x})](\frac{\bar{X}}{\bar{x}})$ . Singh and Sharma (2015) studied the ratio and regression estimator in the presence of non-response and measurement error when non-response occurs on both study variable and auxiliary variable.

Now we will adapt the estimators  $t_1, t_2, \dots, t_{10}$ , and study it in the simultaneous presence of measurement error and non-response.

### 4. Adapted estimators

We have adapted estimators  $t_1, t_2, \dots, t_{10}$  to study them in the presence of non-response and measurement error. Here, we redefine them in two cases and will further investigate.

#### 4.1. Case-1

When non-response occurs only on study variable then  $t_1, t_2, \dots, t_{10}$  can be redefined as

$$1 \quad t_{11} = k_1 \bar{y}^*, k_1 \text{ is constant.}$$

$$2 \quad t_{12} = \bar{y}^* \left( \frac{\bar{X}}{\bar{x}} \right)$$

$$3 \quad t_{13} = \bar{y}^* \left( \frac{\bar{x}}{\bar{X}} \right)$$

$$4 \quad t_{14} = \bar{y}^* + b_1 (\bar{X} - \bar{x}), b_1 \text{ is constant.}$$

$$5 \quad t_{15} = \bar{y}^* \left( \frac{\bar{X}}{\bar{x}} \right)^{\alpha_1}, \alpha_1 \text{ is constant.}$$

$$6 \quad t_{16} = k_{11} \bar{y}^* + k_{12} (\bar{X} - \bar{x}), k_{11}, k_{12} \text{ are constants.}$$

$$7 \quad t_{17} = \bar{y}^* \exp \left( \frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}} \right)$$

$$8 \quad t_{18} = \bar{y}^* \exp \left( \frac{\bar{x} - \bar{X}}{\bar{x} + \bar{X}} \right)$$

$$9 \quad t_{19} = \bar{y}^* \left[ a_1 \left( \frac{\bar{X}}{\bar{x}} \right) + (1 - a_1) \left( \frac{\bar{x}}{\bar{X}} \right) \right], a_1 \text{ is constant.}$$

$$10 \quad t_{20} = [\bar{y}^* + d_1 (\bar{X} - \bar{x})] \left( \frac{\bar{X}}{\bar{x}} \right), d_1 \text{ is constant.}$$

with constants to be determined for minimum MSE.

#### 4.2. Case-2

When non-response occurs on both study and auxiliary variable then  $t_1, t_2, \dots, t_{10}$  can be redefined as

$$1 \quad t_{21} = k_2 \bar{y}^*, k_2 \text{ is constant.}$$

$$2 \quad t_{22} = \bar{y}^* \left( \frac{\bar{X}}{\bar{x}^*} \right)$$

$$3 \quad t_{23} = \bar{y}^* \left( \frac{\bar{x}^*}{\bar{X}} \right)$$

$$4 \quad t_{24} = \bar{y}^* + b_2 (\bar{X} - \bar{x}^*), b_2 \text{ is constant.}$$

$$5 \quad t_{25} = \bar{y}^* \left( \frac{\bar{X}}{\bar{x}^*} \right)^{\alpha_2}, \alpha_2 \text{ is constant.}$$

$$6 \quad t_{26} = k_{21} \bar{y}^* + k_{22} (\bar{X} - \bar{x}^*), k_{21} \text{ and } k_{22} \text{ are constants.}$$

$$7 \quad t_{27} = \bar{y}^* \exp \left( \frac{\bar{X} - \bar{x}^*}{\bar{X} + \bar{x}^*} \right)$$

$$8 \quad t_{28} = \bar{y}^* \exp \left( \frac{\bar{x}^* - \bar{X}}{\bar{x}^* + \bar{X}} \right)$$

$$9 \quad t_{29} = \bar{y}^* \left[ a_2 \left( \frac{\bar{X}}{\bar{x}^*} \right) + (1 - a_2) \left( \frac{\bar{x}^*}{\bar{X}} \right) \right], a_2 \text{ is constant.}$$

$$10 \quad t_{30} = [\bar{y}^* + d_2 (\bar{X} - \bar{x}^*)] \left( \frac{\bar{X}}{\bar{x}^*} \right), d_2 \text{ is constant.}$$

with constants to be determined for minimum MSE.

### 5. Bias and Mean square error

We derive the bias and mean square error (MSE) of the estimators using the following terms.

$$U^* = y_i^* - Y_i^* \text{ and } \omega_Y^* = \frac{1}{\sqrt{n}} \sum_{i=1}^n (Y_i^* - \bar{Y}), \omega_U^* = \frac{1}{\sqrt{n}} \sum_{i=1}^n U_i^*$$

Adding  $\omega_Y^*$  and  $\omega_U^*$  and dividing both side by  $\sqrt{n}$ , we have

$$\frac{\omega_Y^* + \omega_U^*}{\sqrt{n}} = \frac{1}{n} \sum_{i=1}^n [(Y_i^* - \bar{Y}) + U_i^*] \text{ which is } \frac{\omega_Y^* + \omega_U^*}{\sqrt{n}} = \frac{1}{n} \sum_{i=1}^n y_i^* - \bar{Y}$$

So,

$$\bar{y}^* = \bar{Y} + \xi_Y^*; \quad \text{where } \xi_Y^* = \frac{\omega_Y^* + \omega_U^*}{\sqrt{n}}. \tag{1}$$

Similarly, for  $\omega_X = \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \bar{X})$  and  $\omega_V = \frac{1}{\sqrt{n}} \sum_{i=1}^n V_i$ , we have

$$\bar{x} = \bar{X} + \xi_X; \quad \text{where } \xi_X = \frac{\omega_X + \omega_V}{\sqrt{n}}. \tag{2}$$

Again, for  $V^* = x_i^* - X_i^*$  and  $\omega_X^* = \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i^* - \bar{X})$ ,  $\omega_V^* = \frac{1}{\sqrt{n}} \sum_{i=1}^n V_i^*$ , we have

$$\bar{x}^* = \bar{X} + \xi_X^*; \quad \text{where } \xi_X^* = \frac{\omega_X^* + \omega_V^*}{\sqrt{n}}. \tag{3}$$

Expected values of errors are

$$E(\xi_Y^{*2}) = A_{MSE} + A_{ME} = A, \tag{4}$$

where  $A_{MSE} = \lambda S_Y^2 + \theta S_{Y(2)}^2$  and  $A_{ME} = \lambda S_U^2 + \theta S_{U(2)}^2$ .

$$E(\xi_X^2) = B_{MSE} + B_{ME} = B, \tag{5}$$

where  $B_{MSE} = \lambda S_X^2$  and  $A_{ME} = \lambda S_V^2$ .

$$E(\xi_Y^* \xi_X) = \lambda \rho S_Y S_X = C, \tag{6}$$

$$E(\xi_X^{*2}) = D_{MSE} + D_{ME} = D, \tag{7}$$

where  $D_{MSE} = \lambda S_X^2 + \theta S_{X(2)}^2$  and  $D_{ME} = \lambda S_V^2 + \theta S_{V(2)}^2$ .

$$E(\xi_Y^* \xi_X^*) = \lambda \rho S_Y S_X + \theta \rho_2 S_{Y(2)} S_{X(2)} = E, \tag{8}$$

and

$$E(\xi_Y^*) = E(\xi_X) = E(\xi_X^*) = E(\xi_U) = E(\xi_V) = 0. \tag{9}$$

Now, using these values we derive the bias and MSE for all the estimators.

## 5.1. Case-1

### 5.1.1 $t_{11} = k_1 \bar{y}^*$

Using equation (1), express  $t_{11}$  in terms of error as  $t_{11} = k_1(\bar{Y} + \xi_Y^*)$ , so

$$t_{11} - \bar{Y} = (k_1 - 1)\bar{Y} + k_1 \xi_Y^*. \quad (10)$$

Taking expectation on both sides of equation (10), we get

$$\text{Bias}(t_{11}) = (k_1 - 1)\bar{Y}. \quad (11)$$

Now squaring both sides of equation (10), we have

$$(t_{11} - \bar{Y})^2 = (k_1 - 1)^2 \bar{Y}^2 + k_1^2 \xi_Y^{*2} + 2k_1(k_1 - 1)\bar{Y}\xi_Y^*, \quad (12)$$

taking expectation to equation (12), we get

$$\text{MSE}(t_{11}) = (k_1 - 1)^2 \bar{Y}^2 + k_1^2 A. \quad (13)$$

Minimizing  $\text{MSE}(t_{11})$  for  $k_1$ , we get the optimum value of  $k_1$  as  $k_1^o = \frac{\bar{Y}^2}{\bar{Y}^2 + A}$ . Now, putting optimum value of  $k_1$  in equation (13), we get minimum  $\text{MSE}$  of  $t_{11}$ .

$$\text{MSE}_{\min}(t_{11}) = \frac{A\bar{Y}^2}{A + \bar{Y}^2}. \quad (14)$$

### 5.1.2 $t_{12} = \bar{y}^*\left(\frac{\bar{X}}{\bar{x}}\right)$

Expanding  $t_{12}$  using equation (1) and (2), we have  $t_{12} = (\bar{Y} + \xi_Y^*)\frac{\bar{X}}{(\bar{X} + \xi_X)}$  or  $t_{12} = (\bar{Y} + \xi_Y^*)(1 + \frac{\xi_X}{\bar{X}})^{-1}$ .

Assuming  $|\xi| < 1$ , expanding series in the right side and terminating the terms having  $\xi$ 's degree greater than two, we have

$$t_{12} = (\bar{Y} + \xi_Y^*)\left(1 - \frac{\xi_X}{\bar{X}} + \frac{\xi_X^2}{\bar{X}^2}\right),$$

on simplifying, we get

$$t_{12} - \bar{Y} = \xi_Y^* - R\xi_X + \frac{R\xi_X^2}{\bar{X}} - \frac{\xi_Y^*\xi_X}{\bar{X}}, \quad (15)$$

where  $R = \frac{\bar{Y}}{\bar{X}}$ .

Taking expectation on both sides of equation (15), we get

$$\text{Bias}(t_{12}) = \frac{RB - C}{\bar{X}}. \quad (16)$$

Squaring equation (15) and terminating terms with  $\xi$ 's degree greater than two and simplifying, we get

$$(t_{12} - \bar{Y})^2 = \xi_Y^{*2} - R^2 \xi_X^2 - 2R \xi_Y^* \xi_X, \tag{17}$$

taking expectation to equation (17), we get

$$MSE(t_{12}) = A + R^2 B - 2RC. \tag{18}$$

**5.1.3**  $t_{13} = \bar{y}^* \left( \frac{\bar{X}}{\bar{X}} \right)$

Proceeding as in 5.1.2, we get the bias and MSE of  $t_{13}$  as

$$Bias(t_{13}) = \frac{C}{\bar{X}}, \tag{19}$$

and

$$MSE(t_{13}) = A + R^2 B + 2RC. \tag{20}$$

**5.1.4**  $t_{14} = \bar{y}^* + b_1(\bar{X} - \bar{x})$

Using equation (1) and (2) expanding  $t_{14}$ , we get  $t_{14} = \bar{Y} + \xi_Y^* - b_1 \xi_X$  so we have

$$t_{14} - \bar{Y} = \xi_Y^* - b_1 \xi_X. \tag{21}$$

On taking expectation to equation (21), we get

$$Bias(t_{14}) = 0. \tag{22}$$

Squaring both sides of equation (21) and taking expectation, we get

$$MSE(t_{14}) = A + b_1^2 B - 2b_1 C. \tag{23}$$

Minimizing  $MSE(t_{14})$  for  $b_1$ , the optimum value of  $b_1$  is  $b_1^o = \frac{C}{B}$ .

Using optimum value of  $b_1$  in equation (23), we get

$$MSE_{min}(t_{14}) = A - \frac{C^2}{B}. \tag{24}$$

**5.1.5**  $t_{15} = \bar{y}^* \left( \frac{\bar{X}}{\bar{X}} \right)^{\alpha_1}$

Proceeding as in 5.1.2, we get the bias and MSE of  $t_{15}$  as

$$Bias(t_{15}) = \frac{\alpha_1(\alpha_1 + 1)RB - 2\alpha_1 C}{2\bar{X}}, \tag{25}$$

and

$$MSE(t_{15}) = A + \alpha_1^2 R^2 B - 2\alpha_1 RC. \tag{26}$$

Minimizing  $MSE(t_{15})$  for  $\alpha_1$ , the optimum value of  $\alpha_1$  is  $\alpha_1^o = \frac{C}{RB}$ .

Putting optimum value of  $\alpha_1$  in equation (26), we get

$$MSE_{min}(t_{15}) = A - \frac{C^2}{B}. \quad (27)$$

### 5.1.6 $t_{16} = k_{11}\bar{y}^* + k_{12}(\bar{X} - \bar{x})$

Expressing  $t_{16}$  in terms of error using equation (1) and (2), we have  $t_{16} = k_{11}(\bar{Y} + \xi_Y^*) + k_{12}(\bar{X} - \bar{X} - \xi_X)$ . On simplifying, we get

$$t_{16} - \bar{Y} = (k_{11} - 1)\bar{Y} + k_{11}\xi_Y^* - k_{12}\xi_X, \quad (28)$$

taking expectation on both sides of equation (28), we get

$$Bias(t_{16}) = (k_{11} - 1)\bar{Y}. \quad (29)$$

Squaring both sides of equation (28), have

$$(t_{16} - \bar{Y})^2 = (k_{11} - 1)^2\bar{Y}^2 + k_{11}^2\xi_Y^{*2} + k_{12}^2\xi_X^2 + 2k_{11}(k_{11} - 1)\bar{Y}\xi_Y^* - 2k_{11}k_{12}\xi_Y^*\xi_X - 2k_{12}(k_{11} - 1)\bar{Y}\xi_X, \quad (30)$$

taking expectation on both sides of equation (30), we get

$$MSE(t_{16}) = (k_{11} - 1)^2\bar{Y}^2 + k_{11}^2A + k_{12}^2B - 2k_{11}k_{12}C. \quad (31)$$

Minimizing  $MSE(t_{16})$  for  $k_{11}$  and  $k_{12}$ , we get the optimum values of  $k_{11}$  and  $k_{12}$  as  $k_{11}^o = \frac{B\bar{Y}^2}{B\bar{Y}^2 + AB - C^2}$  and  $k_{12}^o = \frac{C\bar{Y}^2}{B\bar{Y}^2 + AB - C^2}$ .

Using optimum values of  $k_{11}$  and  $k_{12}$  in equation (31), we get minimum MSE.

$$MSE_{min}(t_{16}) = \frac{\bar{Y}^2(AB - C^2)}{B\bar{Y}^2 + AB - C^2}. \quad (32)$$

### 5.1.7 $t_{17} = \bar{y}^* \exp\left(\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}}\right)$

Expressing  $t_{17}$  in terms of error as  $t_{17} = (\bar{Y} + \xi_Y^*) \exp\left(\frac{\bar{X} - \bar{X} - \xi_X}{\bar{X} + \bar{X} + \xi_X}\right)$ . On simplifying, we get  $t_{17} = (\bar{Y} + \xi_Y^*) \exp\left[-\frac{\xi_X}{2\bar{X}}\left(1 + \frac{\xi_X}{2\bar{X}}\right)^{-1}\right]$ . Expand the series and ignore the terms having  $\xi$ 's degree greater than two. After simplification, we get  $t_{17} = (\bar{Y} + \xi_Y^*)\left(1 - \frac{\xi_X}{2\bar{X}} + \frac{3}{8}\frac{\xi_X^2}{\bar{X}^2}\right)$ . So, we have

$$t_{17} - \bar{Y} = \xi_Y^* - \frac{R\xi_X}{2} + \frac{3}{8}\frac{R\xi_X^2}{\bar{X}} - \frac{\xi_Y^*\xi_X}{2\bar{X}}. \quad (33)$$

Taking expectation on both sides of equation (33), we get

$$Bias(t_{17}) = \frac{3RB - 4C}{8\bar{X}}. \quad (34)$$

Squaring equation (33) on both sides and taking expectation, we get

$$MSE(t_{17}) = A + \frac{R^2B}{4} - RC. \tag{35}$$

**5.1.8**  $t_{18} = \bar{y}^* \exp(\frac{\bar{x}-\bar{X}}{\bar{x}+\bar{X}})$

Proceeding on the lines of 5.1.7, we get the bias and MSE of  $t_{18}$  as

$$Bias(t_{18}) = \frac{4C - RB}{8\bar{X}}, \tag{36}$$

$$MSE(t_{18}) = A + \frac{R^2B}{4} + RC. \tag{37}$$

**5.1.9**  $t_{19} = \bar{y}^* [a_1(\frac{\bar{X}}{\bar{x}}) + (1 - a_1)(\frac{\bar{x}}{\bar{X}})]$

Express  $t_{19}$  in terms of error as  $t_{19} = (\bar{Y} + \xi_Y^*) [a_1(\frac{\bar{X}}{\bar{x} + \xi_X}) + (1 - a_1)(\frac{\bar{x} + \xi_X}{\bar{X}})]$ . After little simplification  $t_{19} = (\bar{Y} + \xi_Y^*) [a_1(1 + \frac{\xi_X}{\bar{X}})^{-1} + (1 - a_1)(1 + \frac{\xi_X}{\bar{X}})]$ . Expand the series and ignore the terms having  $\xi$ 's degree greater than two. On simplification, we get  $t_{19} = (\bar{Y} + \xi_Y^*) [1 + (1 - 2a_1)\frac{\xi_X}{\bar{X}} + \frac{a_1\xi_X^2}{\bar{X}^2}]$ . So, we have

$$t_{19} - \bar{Y} = \xi_Y^* + (1 - 2a_1)R\xi_X + \frac{a_1R\xi_X^2}{\bar{X}} + (1 - 2a_1)\frac{\xi_Y^*\xi_X}{\bar{X}}. \tag{38}$$

Taking expectation on both sides of equation (38), we get

$$Bias(t_{19}) = \frac{a_1RB + (1 - 2a_1)C}{\bar{X}}. \tag{39}$$

Squaring both sides of equation (38) and taking expectation, we get

$$MSE(t_{19}) = A + (1 - 2a_1)^2R^2B + 2R(1 - 2a_1)C. \tag{40}$$

Minimizing  $MSE(t_{19})$  for  $a_1$ , we get the optimum value of  $a_1$  as  $a_1^o = \frac{1}{2} + \frac{C}{2RB}$ .

Putting optimum value of  $a_1$  in equation (40), we get

$$MSE_{min}(t_{19}) = A - \frac{C^2}{B}. \tag{41}$$

**5.1.10**  $t_{20} = [\bar{y}^* + d_1(\bar{X} - \bar{x})](\frac{\bar{X}}{\bar{x}})$

Expressing  $t_{20}$  in terms of error and simplifying, we get

$$t_{20} - \bar{Y} = \xi_Y^* - (R + d_1)\xi_X + \frac{(R + d_1)\xi_X^2}{\bar{X}} - \frac{\xi_Y^*\xi_X}{\bar{X}}, \tag{42}$$

taking expectation on both sides of equation (42), we get

$$\text{Bias}(t_{20}) = \frac{(R + d_1)B - C}{\bar{X}}. \quad (43)$$

Squaring both sides of equation (42) and taking expectation, we get

$$\text{MSE}(t_{20}) = A + (R + d_1)^2 B - 2(R + d_1)C. \quad (44)$$

Minimizing  $\text{MSE}(t_{20})$  for  $d_1$ , we get the optimum value of  $d_1$  as  $d_1^o = \frac{C}{B} - R$ .

Putting optimum value of  $d_1$  in equation (44), we get

$$\text{MSE}_{\min}(t_{20}) = A - \frac{C^2}{B}. \quad (45)$$

The bias and MSEs in the next case can be obtained in similar steps used in Case-1. To save space, only the results are given.

## 5.2. Case-2

### 5.2.1 $t_{21} = k_2 \bar{y}^*$

$$\text{Bias}(t_{21}) = (k_2 - 1)\bar{Y}, \quad (46)$$

$$\text{MSE}(t_{21}) = (k_2 - 1)^2 \bar{Y}^2 + k_2^2 A. \quad (47)$$

Optimum value of  $k_2$  is  $k_2^o = \frac{\bar{Y}^2}{\bar{Y}^2 + A}$ .

$$\text{MSE}_{\min}(t_{21}) = \frac{A\bar{Y}^2}{A + \bar{Y}^2}. \quad (48)$$

### 5.2.2 $t_{22} = \bar{y}^* \left( \frac{\bar{X}}{\bar{y}^*} \right)$

$$\text{Bias}(t_{22}) = \frac{RD - E}{\bar{X}}, \quad (49)$$

$$\text{MSE}(t_{22}) = A + R^2 D - 2RE. \quad (50)$$

### 5.2.3 $t_{23} = \bar{y}^* \left( \frac{\bar{y}^*}{\bar{X}} \right)$

$$\text{Bias}(t_{23}) = \frac{E}{\bar{X}}, \quad (51)$$

$$\text{MSE}(t_{23}) = A + R^2 D + 2RE. \quad (52)$$



**5.2.4**  $t_{24} = \bar{y}^* + b_2(\bar{X} - \bar{x}^*)$

$$Bias(t_{24}) = 0, \tag{53}$$

$$MSE(t_{24}) = A + b_2^2 D - 2b_2 E. \tag{54}$$

Optimum value of  $b_2$  is  $b_2^o = \frac{E}{D}$ .

$$MSE_{min}(t_{24}) = A - \frac{E^2}{D}. \tag{55}$$

**5.2.5**  $t_{25} = \bar{y}^* \left(\frac{\bar{X}}{\bar{x}^*}\right)^{\alpha_2}$

$$Bias(t_{25}) = \frac{\alpha_2(\alpha_2 + 1)RD - 2\alpha_2 E}{2\bar{X}}, \tag{56}$$

$$MSE(t_{25}) = A + \alpha_2^2 R^2 D - 2\alpha_2 RE. \tag{57}$$

Optimum value of  $\alpha_2$  is  $\alpha_2^o = \frac{E}{RD}$ .

$$MSE_{min}(t_{25}) = A - \frac{E^2}{D}. \tag{58}$$

**5.2.6**  $t_{26} = k_{21}\bar{y}^* + k_{22}(\bar{X} - \bar{x}^*)$

$$Bias(t_{26}) = (k_{21} - 1)\bar{Y}, \tag{59}$$

$$MSE(t_{26}) = (k_{21} - 1)^2 \bar{Y}^2 + k_{21}^2 A + k_{22}^2 D - 2k_{21}k_{22}E. \tag{60}$$

Optimum values of  $k_{21}$  and  $k_{22}$  are  $k_{21}^o = \frac{D\bar{Y}^2}{D\bar{Y}^2 + AD - E^2}$  and  $k_{22}^o = \frac{E\bar{Y}^2}{D\bar{Y}^2 + AD - E^2}$ .

$$MSE_{min}(t_{26}) = \frac{\bar{Y}^2(AD - E^2)}{D\bar{Y}^2 + AD - E^2}. \tag{61}$$

**5.2.7**  $t_{27} = \bar{y}^* \exp\left(\frac{\bar{X} - \bar{x}^*}{\bar{X} + \bar{x}^*}\right)$

$$Bias(t_{27}) = \frac{3RD - 4E}{8\bar{X}}, \tag{62}$$

$$MSE(t_{27}) = A + \frac{R^2 D}{4} - RE. \tag{63}$$

**5.2.8**  $t_{28} = \bar{y}^* \exp\left(\frac{\bar{x}^* - \bar{X}}{\bar{x}^* + \bar{X}}\right)$

$$Bias(t_{28}) = \frac{4E - RD}{8\bar{X}}, \tag{64}$$

$$MSE(t_{28}) = A + \frac{R^2 D}{4} + RE. \quad (65)$$

$$5.2.9 \quad t_{29} = \bar{y}^* [a_2 (\frac{\bar{X}}{\bar{y}^*}) + (1 - a_2) (\frac{\bar{x}^*}{\bar{X}})]$$

$$Bias(t_{29}) = \frac{a_1 R D + (1 - 2a_2) E}{\bar{X}}, \quad (66)$$

$$MSE(t_{29}) = A + (1 - 2a_2)^2 R^2 D + 2R(1 - 2a_2)E. \quad (67)$$

Optimum value of  $a_2$  is  $a_2^o = \frac{1}{2} + \frac{E}{2RD}$ .

$$MSE_{min}(t_{29}) = A - \frac{E^2}{D}. \quad (68)$$

$$5.2.10 \quad t_{30} = [\bar{y}^* + d_2 (\bar{X} - \bar{x}^*)] (\frac{\bar{X}}{\bar{y}^*})$$

$$Bias(t_{30}) = \frac{(R + b_2)D - E}{\bar{X}}, \quad (69)$$

$$MSE(t_{30}) = A + (R + b_2)^2 D - 2(R + b_2)E. \quad (70)$$

Optimum value of  $d_2$  is  $d_2^o = \frac{E}{D} - R$ .

$$MSE_{min}(t_{30}) = A - \frac{E^2}{D}. \quad (71)$$

## Note

The optimum MSEs of  $t_{i4}$ ,  $t_{i5}$ ,  $t_{i9}$  and  $t_{j0}$  are equal, where  $i = 1, 2$  and  $j = 2, 3$ .

## 6. Efficiency comparison

In this section, we derive the conditions under which the estimators perform better than the usual estimator  $\bar{y}^*$ . As we know, an estimator  $t$  will be more efficient than  $\bar{y}^*$  whenever the inequality  $var(\bar{y}^*) - MSE(t) > 0$  is satisfied.

The variance of usual estimator  $\bar{y}^*$  in the presence of non-response and measurement error is  $var(\bar{y}^*) = \lambda S_Y^2 + \theta S_{Y(2)}^2 + \lambda S_U^2 + \theta S_{U(2)}^2$ . That is

$$var(\bar{y}^*) = A. \quad (72)$$

### 6.1. Case-1

$$1. \quad MSE(t_{11}) < var(\bar{y}^*) \quad \text{if} \quad \frac{A^2}{A+Y^2} > 0$$

$$2. \quad MSE(t_{12}) < var(\bar{y}^*) \quad \text{if} \quad \frac{C}{B} > \frac{R}{2}$$

3.  $MSE(t_{13}) < var(\bar{y}^*)$  if  $\frac{C}{B} < -\frac{R}{2}$
4.  $MSE(t_{14}) < var(\bar{y}^*)$  if  $\frac{C^2}{B} > 0$
5.  $MSE(t_{15}) < var(\bar{y}^*)$  if  $\frac{C^2}{B} > 0$
6.  $MSE(t_{16}) < var(\bar{y}^*)$  if  $\frac{A^2B-AC^2+\bar{Y}^2C^2}{B\bar{Y}^2+AB-C^2} > 0$
7.  $MSE(t_{17}) < var(\bar{y}^*)$  if  $\frac{C}{B} > \frac{R}{4}$
8.  $MSE(t_{18}) < var(\bar{y}^*)$  if  $\frac{C}{B} < -\frac{R}{4}$
9.  $MSE(t_{19}) < var(\bar{y}^*)$  if  $\frac{C^2}{B} > 0$
10.  $MSE(t_{20}) < var(\bar{y}^*)$  if  $\frac{C^2}{B} > 0$

**6.2. Case-2**

1.  $MSE(t_{21}) < var(\bar{y}^*)$  if  $\frac{A^2}{A+\bar{Y}^2} > 0$
2.  $MSE(t_{22}) < var(\bar{y}^*)$  if  $\frac{E}{D} > \frac{R}{2}$
3.  $MSE(t_{23}) < var(\bar{y}^*)$  if  $\frac{E}{D} < -\frac{R}{2}$
4.  $MSE(t_{24}) < var(\bar{y}^*)$  if  $\frac{E^2}{D} > 0$
5.  $MSE(t_{25}) < var(\bar{y}^*)$  if  $\frac{E^2}{D} > 0$
6.  $MSE(t_{26}) < var(\bar{y}^*)$  if  $\frac{A^2D-AE^2+\bar{Y}^2C^2}{D\bar{Y}^2+AD-E^2} > 0$
7.  $MSE(t_{27}) < var(\bar{y}^*)$  if  $\frac{E}{D} > \frac{R}{4}$
8.  $MSE(t_{28}) < var(\bar{y}^*)$  if  $\frac{E}{D} < -\frac{R}{4}$
9.  $MSE(t_{29}) < var(\bar{y}^*)$  if  $\frac{E^2}{D} > 0$
10.  $MSE(t_{30}) < var(\bar{y}^*)$  if  $\frac{E^2}{D} > 0$

**7. Monte-Carlo Simulation**

For validating the theoretical results in the previous sections, we perform a Monte-Carlo simulation study. We have used the following information to generate the data in R software:  $N = 4000$ ,  $n = 500$ ,  $X = rnorm(N, 4, 7)$ ,  $Y = 1 + 2X + \epsilon$ ,  $\epsilon = rnorm(N, 0, 1)$ ,  $U = rnorm(N, 0, 3)$ ,  $V = rnorm(N, 0, 3)$ . We have checked the performance of estimators for a different response rate. To get output more accurate, we have made 10000 replications to the process.

Percent relative efficiency (PRE) of an estimator  $t$  with respect to  $\bar{y}^*$  is calculated by

$$PRE(t, \bar{y}^*) = \frac{var(\bar{y}^*)}{MSE(t)} \times 100. \tag{73}$$

To get PRE without measurement error, we use MSEs without terms of measurement error. That is,  $var(\bar{y}^*) = \lambda S_Y^2 + \theta S_{Y(2)}^2$  and  $A = A_{MSE} = \lambda S_Y^2 + \theta S_{Y(2)}^2$ ,  $B = B_{MSE} = \lambda S_X^2$ ,  $D = D_{MSE} = \lambda S_X^2 + \theta S_{X(2)}^2$  are used in expressions of MSEs of estimators.

We have compared the estimators using PREs in Table 1 and Table 2. From the definition of PRE in equation (73), higher PRE of an estimator means smaller MSE.

Table 1: PREs of estimators with respect to  $\bar{y}^*$  in Case-1

$N_1$	$N_2$	Estimator	$PRE(., \bar{y}^*)$ without measurement error					$PRE(., \bar{y}^*)$ with measurement error				
			1/k					1/k				
			1/2	1/3	1/4	1/5	1/10	1/2	1/3	1/4	1/5	1/10
500		$\bar{y}^*$	100	100	100	100	100	100	100	100	100	100
		$t_{11}$	100.48	100.54	100.60	100.66	100.97	100.50	100.57	100.63	100.69	101.01
		$t_{12}$	699.90	420.18	318.36	265.68	175.09	267.41	225.51	200.39	183.64	145.61
		$t_{13}$	24.62	26.86	28.98	30.98	39.49	24.26	26.49	28.59	30.57	39.03
		$t_{14}$	773.62	442.70	329.80	272.86	177.20	337.22	266.94	228.79	204.83	154.31
		$t_{15}$	773.62	442.70	329.80	272.86	177.20	337.22	266.94	228.79	204.83	154.31
		$t_{16}$	774.11	443.25	330.41	273.53	178.17	337.72	267.51	229.42	205.53	155.33
		$t_{17}$	337.94	267.35	229.06	205.03	154.39	266.60	225.00	200.02	183.37	145.49
		$t_{18}$	44.34	47.26	49.88	52.26	61.42	44.47	47.39	50.01	52.39	61.54
		$t_{19}$	773.62	442.70	329.80	272.86	177.20	337.22	266.94	228.79	204.83	154.31
		$t_{20}$	773.62	442.70	329.80	272.86	177.20	337.22	266.94	228.79	204.83	154.31
		1000		$\bar{y}^*$	100	100	100	100	100	100	100	100
$t_{11}$	100.54			100.66	100.79	100.91	101.52	100.57	100.69	100.82	100.95	101.59
$t_{12}$	419.76			265.45	211.59	184.19	137.79	225.43	183.57	162.66	150.66	125.05
$t_{13}$	26.87			30.99	34.67	37.98	50.51	26.49	30.58	34.24	37.53	50.03
$t_{14}$	442.22			272.61	215.41	186.69	138.62	266.81	204.72	176.32	160.03	129.04
$t_{15}$	442.22			272.61	215.41	186.69	138.62	266.81	204.72	176.32	160.03	129.04
$t_{16}$	442.77			273.28	216.20	187.60	140.14	267.38	205.42	177.14	160.99	130.63
$t_{17}$	267.21			204.92	176.44	160.12	129.08	224.92	183.29	162.47	149.98	124.99
$t_{18}$	47.26			52.28	56.42	59.90	71.34	47.39	52.40	56.54	60.02	71.45
$t_{19}$	442.22			272.61	215.41	186.69	138.62	266.81	204.72	176.32	160.03	129.04
$t_{20}$	442.22			272.61	215.41	186.69	138.62	266.81	204.72	176.32	160.03	129.04
1500				$\bar{y}^*$	100	100	100	100	100	100	100	100
		$t_{11}$	100.60	100.79	100.97	101.15	102.06	100.63	100.82	101.01	101.20	102.16
		$t_{12}$	317.98	211.55	174.95	156.43	125.25	200.28	162.64	145.54	135.78	117.26
		$t_{13}$	28.99	34.68	39.52	43.69	58.13	28.60	34.24	39.06	43.21	57.66
		$t_{14}$	329.38	215.36	177.06	157.85	125.75	228.63	176.29	154.23	142.06	119.82
		$t_{15}$	329.38	215.36	177.06	157.85	125.75	228.63	176.29	154.23	142.06	119.82
		$t_{16}$	329.99	216.16	178.03	159.01	127.82	229.27	177.12	155.24	143.27	121.99
		$t_{17}$	228.89	176.41	154.30	142.11	119.84	199.92	162.45	145.42	135.69	117.23
		$t_{18}$	49.90	56.42	61.44	65.43	77.20	50.03	56.55	61.57	65.54	77.29
		$t_{19}$	329.38	215.36	177.06	157.85	125.75	228.63	176.29	154.23	142.06	119.82
		$t_{20}$	329.38	215.36	177.06	157.85	125.75	228.63	176.29	154.23	142.06	119.82

Table 2: PREs of estimators with respect to  $\bar{y}^*$  in Case-2

$N_1$	$N_2$	Estimator	$PRE(., \bar{y}^*)$ without measurement error					$PRE(., \bar{y}^*)$ with measurement error				
			$1/k$					$1/k$				
			1/2	1/3	1/4	1/5	1/10	1/2	1/3	1/4	1/5	1/10
3500	500	$\bar{y}^*$	100	100	100	100	100	100	100	100	100	100
		$t_{21}$	100.48	100.54	100.60	100.66	100.97	100.50	100.57	100.63	100.69	101.01
		$t_{22}$	4845.82	4845.35	4844.98	4844.68	4843.73	351.22	351.15	351.11	351.07	350.94
		$t_{23}$	22.23	22.23	22.23	22.23	22.23	21.89	21.89	21.89	21.89	21.89
		$t_{24}$	19691.95	19684.57	196778.67	19673.84	19658.76	509.55	509.47	509.41	509.36	509.19
		$t_{25}$	19691.95	19684.57	196778.67	19673.84	19658.76	509.55	509.47	509.41	509.36	509.19
		$t_{26}$	19692.44	19685.12	19679.28	19674.51	19659.73	510.06	510.05	510.05	510.06	510.21
		$t_{27}$	511.53	511.52	511.52	511.52	511.52	399.68	399.66	399.64	349.63	399.59
		$t_{28}$	41.08	41.08	41.08	41.08	41.08	41.20	41.20	41.20	41.20	41.20
		$t_{29}$	19691.95	19684.57	196778.67	19673.84	19658.76	509.55	509.47	509.41	509.36	509.19
		$t_{30}$	19691.95	19684.57	196778.67	19673.84	19658.76	509.55	509.47	509.41	509.36	509.19
		3000	1000	$\bar{y}^*$	100	100	100	100	100	100	100	100
$t_{21}$	100.54			100.66	100.79	100.91	101.52	100.57	100.69	100.82	100.95	101.59
$t_{22}$	4845.45			4845.65	4845.94	4845.29	4844.90	351.31	351.32	351.33	351.33	
$t_{23}$	22.23			22.23	22.23	22.23	22.23	21.89	21.89	21.89	21.89	21.89
$t_{24}$	19692.31			19686.50	19682.48	19679.54	19671.88	509.67	509.68	509.68	509.69	509.70
$t_{25}$	19692.31			19686.50	19682.48	19679.54	19671.88	509.67	509.68	509.68	509.69	509.70
$t_{26}$	19692.86			19687.17	19683.27	19680.45	19673.40	510.24	510.38	510.51	510.64	511.29
$t_{27}$	511.53			511.53	511.53	511.53	511.51	349.71	349.70	349.70	349.70	349.70
$t_{28}$	41.08			41.08	41.08	41.08	41.08	41.20	41.20	41.20	41.20	41.20
$t_{29}$	19692.31			19686.50	19682.48	19679.54	19671.88	509.67	509.68	509.68	509.69	509.70
$t_{30}$	19692.31			19686.50	19682.48	19679.54	19671.88	509.67	509.68	509.68	509.69	509.70
2500	1500			$\bar{y}^*$	100	100	100	100	100	100	100	100
		$t_{21}$	100.60	100.79	100.97	101.15	102.06	100.63	100.82	101.01	101.20	102.16
		$t_{22}$	4846.28	4846.21	4846.17	4846.14	4846.07	351.36	351.40	351.42	351.44	351.47
		$t_{23}$	22.23	22.23	22.23	22.23	22.23	21.89	21.89	21.89	21.89	21.90
		$t_{24}$	19693.88	19689.81	19687.26	19685.52	19681.43	509.74	509.78	509.80	509.82	509.86
		$t_{25}$	19693.88	19689.81	19687.26	19685.52	19681.43	509.74	509.78	509.80	509.82	509.86
		$t_{26}$	19694.49	19690.60	19688.24	19686.68	19683.50	510.37	510.61	510.82	511.03	512.03
		$t_{27}$	511.52	511.52	511.51	511.51	511.51	349.72	349.73	349.74	349.74	349.75
		$t_{28}$	41.08	41.08	41.08	41.08	41.08	41.20	41.20	41.20	41.20	41.20
		$t_{29}$	19693.88	19689.81	19687.26	19685.52	19681.43	509.74	509.78	509.80	509.82	509.86
		$t_{30}$	19693.88	19689.81	19687.26	19685.52	19681.43	509.74	509.78	509.80	509.82	509.86

From Table 1 and 2, it is concluded that:

1. Searls (1964) estimator  $t_{i2}$  has minute advantage over usual estimator  $\bar{y}^*$ .
2. PREs of estimators  $t_{i5}$ ,  $t_{i9}$  and  $t_{j0}$  are equal to the PRE of regression estimator  $t_{i4}$ , as their optimum MSEs are the same.
3. Rao (1991) estimator  $t_{i6}$  perform slightly better than regression estimator.
4. In all the estimators, Rao (1991) estimator  $t_{i6}$  performs best in terms of having highest PREs.

Here  $i = 1, 2$  and  $j = 2, 3$ .

## 8. Conclusion

We have considered ten estimators of population mean and studied them in the context of non-response and measurement error. We have obtained the expressions for bias and MSE for all the estimators in various cases. It is noted that optimum MSEs of Srivastava (1967) estimator  $t_{i5}$ , Singh and Espejo (2003) estimator  $t_{i9}$  and Kadilar and Cingi (2004) estimator  $t_{j0}$  are the same which is equal to the MSE of the regression estimator  $t_{i4}$  within the same sampling strategy  $i$ ,  $i = 1, 2$ ;  $j = 2, 3$ . This is also verified in the simulation study. It is also worth to mention that Rao (1991) difference estimator  $t_{i6}$  performs better than other estimators, although its efficiency over the regression estimator is very minute.

## References

- Azeem, M., Hanif, M., (2016). Joint influence of measurement error and non-response on estimation of population mean. *Commun Statist. Theory Methods*, 14(1), 12. DOI: 10.1080/03610926.2015.1026992.
- Bahl, S., Tuteja, R. K., (1991). Ratio and product type exponential estimators. *Journal of Information and Optimization Sciences*, 12, pp. 159–164.
- Cochran, W. G., (1940). The estimation of the yields of cereal experiments by sampling for the ratio gain to total produce. *J. Agric. Soc.*, 30, pp. 262–275.
- Cochran, W. G., (1977). Sampling techniques. *New York: Wiley*.
- Diane, G., Giordan, M., (2012). Finite population variance estimation in presence of measurement errors. *Communications in Statistics - Theory and Methods*, 41(23), pp. 4302–4314. <https://doi.org/10.1080/03610926.2011.573165>.
- Fuller, W. A., (1987), Measurement Error Models, *New York: Wiley*.

- Gregoire, T. G., Salas, C., (2009). Ratio estimation with measurement error in the auxiliary variate. *Biometrics*, 65(2), 590—598.  
DOI:10.1111/j.1541-0420.2008.01110.x.
- Hansen, M. H., Hurwitz, W. N., (1946). The problem of non-response in sample surveys. *J Amer Statist Assoc*, 41, pp. 517–529.
- Kumar, S., Bhogal, S., Nataraja, N. S., (2015). Estimation of population mean in the presence of non-response and measurement error. *Revista Colombiana de Estadística*, 38(1), pp. 145—61. DOI:10.15446/rce.v38n1.48807.
- Kumar, S., Bhogal, S., (2018). Study on Non Response and Measurement Error using Double Sampling Scheme. *J. Stat. Appl. Pro. Lett.*, 5(1), pp. 43–52.
- Kreuter, F., Olson, K., Wagner, J., et al., (2010). Using proxy measure and correlates of survey outcomes to adjust for non-response-examples from multiple surveys. *J Royal Statist Soc Ser A*, 173(3), pp. 1–21.
- Khan, M., Shabbir, J., Hussain, Z., et al., (2014). A Class of Estimators for Finite Population Mean in Double Sampling under Nonresponse Using Fractional Raw Moments. *Journal of Applied Mathematics, Volume*. <http://dx.doi.org/10.1155/2014/282065>.
- Kadilar, C., Cingi, H., (2004). Ratio estimators in simple random sampling. *Applied Mathematics and Computation*, 151(3), pp. 893—902. DOI 10.1080/03610926.2019.1682167.
- Khare, B. B., Sinha, R. R., (2019). Estimation of product of two population means by multi-auxiliary characters under double sampling the non-respondent. *STATISTICS IN TRANSITION new series*, 20(3), pp. 81—95. DOI: 10.21307/stattrans-2019-025.
- Khare, B. B., Srivastava, S., (1993). Estimation of population mean using auxiliary character in presence of non-response. *National Academy of Science and Letters, India*, 16(3), pp. 111–114.
- Luengo, A. V. G., (2016). Ratio-cum-product estimation in presence of non-response in successive sampling. *JAMSI*, 12(1), pp. 55–83.  
<https://doi.org/10.1515/jamsi-2016-0005>.
- Murthy, M. N., (1964). Product method of estimation. *Sankhya A*, 26, pp. 69—74.
- Okafor, F. C., Lee, H., (2000). Double sampling for ratio and regression estimation with sub-sampling the non-respondents. *Survey Methodology*, 26, pp. 183—188.
- Pandey, A. K., Usman, M., Singh, G. N., (2021). Optimality of ratio and regression type estimators using dual of auxiliary variable under non response, *Alexandria Engineer-*

- ing Journal*, 60(5), pp. 4461–4471. <https://doi.org/10.1016/j.aej.2021.03.031>.
- Rao, P. S. R. S., (1986). Ratio estimation with sub-sampling the non-respondents. *Survey Methodology*, 12, pp. 217–230.
- Rao, T. J., (1991). On certain methods of improving ratio and regression estimators, *Communications in Statistics-Theory and Methods*, 20(10), pp. 3325–3340.
- Sharma, V., Kumar, S., (2020). Estimation of population mean using transformed auxiliary variable and non-response. *Revista Investigacio Operacional*, 41(3), pp. 438–444.
- Sinha, R. R., Dhingra, H., Thakur, P., (2022). Estimation of Ratio of Two Means Using Regression-Cum-Exponential Estimators in the Presence of Non-response. *Proc. Natl. Acad. Sci., India, Sect. A Phys. Sci.*, 92, pp. 57–64. <https://doi.org/10.1007/s40010-020-00690-0>.
- Shalabh., (1997). Ratio method of estimation in the presence of measurement errors, *Indian Journal of Agricultural Statistics*, 50(2), pp. 150–155.
- Srivastava, A. K., Shalabh., (2001). Effect of measurement errors on the regression method of estimation in survey sampling. *Journal of Statistical Research*, 35(2), pp. 35–44.
- Singh, R. S., Sharma, P., (2015). Method of Estimation in the Presence of Non-response and Measurement Errors Simultaneously. *Journal of Modern Applied Statistical Methods*, 14(1), pp. 107-121. DOI: 10.22237/jmasm/1430453460.
- Singh, P., Singh, R., Bouza, C. N., (2018). Effect of measurement error and non-response on estimation of population mean. *Revista Investigacion Operacional*, 39(1), pp. 108–120.
- Searls, D. T., (1964). The Utilization of a Known Coefficient of Variation in the Estimation Procedure. *Journal of the American Statistical Association*, 59, pp. 1225–1226.
- Srivastava, S. K., (1967). An estimator using auxiliary information in sample surveys, *Calcutta Statistical Association Bulletin*, 16, pp. 121–132.
- Singh, H. P., Espejo, M. R., (2003). On Linear Regression and Ratio-product Estimation of a Finite Population Mean. *The Statistician*, 52(1), pp. 59–67.
- Tiwari, K. K., Bhougal, S., Kumar, S., Rather, K. U. I., (2022). Using Randomized Response to Estimate the Population Mean of a Sensitive Variable under the Influence of Measurement Error. *Journal of Statistical Theory and Practice*, 16(2), pp. 1-11. <https://doi.org/10.1007/s42519-022-00251-1>.



- Tiwari, K. K., Bhougal, S., Kumar, S., Onyango, R., (2022). Assessing the Effect of Nonresponse and Measurement Error Using a Novel Class of Efficient Estimators. *Journal of Mathematics*, Article ID 4946265. DOI: 10.1155/2022/4946265.
- Tiwari, K. K., Bhougal, S., Kumar, S., (2023). A General Class of Estimators in the Presence of Non-response and Measurement Error, *Journal of Statistics Application & Probability Letters*, 10(1), pp. 13-33. <http://dx.doi.org/10.18576/jsapl/100102>.
- Zahid, E., Shabbir, J., Gupta, S., Onyango, R., Saeed, S., (2022). A generalized class of estimators for sensitive variable in the presence of measurement error and non-response. *PLoS ONE* 17(1), e0261561. <https://doi.org/10.1371/journal.pone.0261561>.

## Appendix

Here, we have to prove the equations (4), (5), (6), (7) and (8).

Using equation (1), we have

$$\xi_Y^* = \bar{y}^* - \bar{Y}.$$

Squaring and taking expectation on both sides, we have

$$E(\xi_Y^{*2}) = E(\bar{y}^* - \bar{Y})^2,$$

that is,

$$E(\xi_Y^{*2}) = V(\bar{y}^*). \quad (74)$$

here,  $V$  represents variance.

Since  $\bar{y}^* = \bar{y}_t^* + \bar{U}^*$ , so  $V(\bar{y}^*) = V(\bar{y}_t^*) + V(\bar{U}^*) + Cov(\bar{y}_t^*, \bar{U}^*)$ . As  $y$  and  $U$  are independent, so  $Cov(\bar{y}_t^*, \bar{U}^*) = 0$ . We have,

$$V(\bar{y}^*) = V(\bar{y}_t^*) + V(\bar{U}^*). \quad (75)$$

Now, we have to derive  $V(\bar{y}_t^*)$ .

$$\bar{y}_t^* = \left(\frac{n_1}{n}\right)\bar{y}_{n_1} + \left(\frac{n_2}{n}\right)\bar{y}_r,$$

so,

$$V(\bar{y}_t^*) = V_1[E_2(\bar{y}_t^*|n_1, n_2)] + E_1[V_2(\bar{y}_t^*|n_1, n_2)]. \quad (76)$$

Considering the first part of (76), we have

$$\begin{aligned} V_1[E_2(\bar{y}_t^*|n_1, n_2)] &= V_1 \left[ E_2 \left\{ \left( \frac{n_1}{n}\bar{y}_{n_1} + \frac{n_2}{n}\bar{y}_r \right) | n_1, n_2 \right\} \right] \\ &= V_1 \left[ \frac{n_1}{n} E_2(\bar{y}_{n_1}) | n_1 + \frac{n_2}{n} E_2(\bar{y}_r) | n_2 \right] \\ &= V_1 \left[ \frac{n_1}{n} \bar{y} + \frac{n_2}{n} \bar{y} \right] \\ &= V_1(\bar{y}) \\ &= \lambda S_Y^2. \end{aligned} \quad (77)$$

Considering the second part of equation (76), we have

$$\begin{aligned} E_1[V_2(\bar{y}_t^*|n_1, n_2)] &= E_1 \left[ V_2 \left\{ \left( \frac{n_1}{n}\bar{y}_{n_1} + \frac{n_2}{n}\bar{y}_r \right) | n_1, n_2 \right\} \right] \\ &= E_1 \left[ V_2 \left\{ \frac{n_1}{n}\bar{y}_{n_1} | n_1 \right\} + V_2 \left\{ \frac{n_2}{n}\bar{y}_r | n_2 \right\} \right] \\ &= E_1 \left[ \frac{n_2^2}{n^2} \left( \frac{1}{r} - \frac{1}{n_2} \right) S_r^2 \right] \\ &= \frac{n_2^2}{n^2} \left( \frac{1}{r} - \frac{1}{n_2} \right) E_1(S_r^2) \\ &= \frac{n_2}{n^2} \left( \frac{n_2}{r} - 1 \right) S_{Y(2)}^2 \\ &= \frac{W_2(k-1)}{n} S_{Y(2)}^2 \\ &= \theta S_{Y(2)}^2. \end{aligned} \quad (78)$$

Using equations (77), (78) in (76), we have

$$V(\bar{y}_t^*) = \lambda S_Y^2 + \theta S_{Y(2)}^2. \tag{79}$$

Similarly, we can derive

$$V(\bar{x}_t^*) = \lambda S_X^2 + \theta S_{X(2)}^2, \tag{80}$$

$$V(\bar{U}^*) = \lambda S_U^2 + \theta S_{U(2)}^2, \tag{81}$$

$$V(\bar{V}^*) = \lambda S_V^2 + \theta S_{V(2)}^2. \tag{82}$$

So, using equations (79), (81) in (75), we have

$$V(\bar{y}^*) = \lambda (S_Y^2 + S_U^2) + \theta (S_{Y(2)}^2 + S_{Y(2)}^2). \tag{83}$$

From equation (74) and (83), we have

$$E(\xi_Y^{*2}) = \lambda (S_Y^2 + S_U^2) + \theta (S_{Y(2)}^2 + S_{Y(2)}^2). \tag{84}$$

Which completes the proof of equation (4).

Similarly, we can show that

$$E(\xi_X^{*2}) = \lambda \{S_X^2 + S_V^2\} + \theta \{S_{X(2)}^2 + S_{V(2)}^2\}, \tag{85}$$

and

$$E(\xi_Z^{*2}) = \lambda \{S_X^2 + S_V^2\}. \tag{86}$$

Now, using equation (1) and (3), we have

$$\xi_Y^* \xi_X^* = (\bar{y}^* - \bar{Y})(\bar{x}^* - \bar{X}),$$

Taking expectation on both sides, we have

$$E(\xi_Y^* \xi_X^*) = Cov(\bar{y}^*, \bar{x}^*). \tag{87}$$

Now,

$$Cov(\bar{y}^*, \bar{x}^*) = E_1 [Cov_2\{(\bar{y}^*, \bar{x}^*)|n_1, n_2\}] + Cov_1 [E_2\{\bar{x}^*|n_1, n_2\}, E_2\{\bar{y}^*|n_1, n_2\}], \tag{88}$$

considering the second part,

$$\begin{aligned} Cov_1 [E_2\{\bar{x}^*|n_1, n_2\}, E_2\{\bar{y}^*|n_1, n_2\}] &= Cov_1 [E_2\{(w_1 \bar{x}_{n_1} + w_2 \bar{x}_r)|n_1, n_2\}, \\ &\quad E_2\{(w_1 \bar{y}_{n_1} + w_2 \bar{y}_r)|n_1, n_2\}] \\ &= Cov_1 [\{w_1 \bar{x} + w_2 \bar{x}\}, \{w_1 \bar{y} + w_2 \bar{y}\}] \\ &= Cov(\bar{x}, \bar{y}) \\ &= \lambda \rho S_Y S_X. \end{aligned} \tag{89}$$

Now, the first part of equation (88)

$$\begin{aligned}
 E_1[\text{Cov}_2\{\bar{y}^*, \bar{x}^*\}|n_1, n_2\}] &= E_1[\text{Cov}_2\{(w_1\bar{x}_{n_1} + w_2\bar{x}_r)|n_1, n_2, (w_1\bar{y}_{n_1} + w_2\bar{y}_r)|n_1, n_2\}] \\
 &= E_1[\text{Cov}_2\{(w_1\bar{x}_{n_1}, w_1\bar{y}_{n_1})|n_1, n_2\} + \text{Cov}_2\{(w_1\bar{x}_{n_1}, w_2\bar{y}_r)|n_1, n_2\} \\
 &\quad + \text{Cov}_2\{(w_2\bar{x}_r, w_1\bar{y}_{n_1})|n_1, n_2\} + \text{Cov}_2\{(w_2\bar{x}_r, w_2\bar{y}_r)|n_1, n_2\}] \\
 &= E_1[(w_2^2 \text{Cov}_2\{\bar{x}_r, \bar{y}_r\}|n_2\}] \\
 &= w_2^2 \left( \frac{1}{r} - \frac{1}{n_2} \right) E_1(S_{r_{YX(2)}}) \\
 &= \frac{n_2}{n^2} \left( \frac{n_2}{r} - 1 \right) S_{YX(2)} \\
 &= \theta S_{YX(2)} \\
 &= \theta \rho_2 S_{Y(2)} S_{X(2)}. \tag{90}
 \end{aligned}$$

Using equations (89), (90) in (88), we have

$$\text{Cov}(\bar{y}^*, \bar{x}^*) = \lambda \rho S_Y S_X + \theta \rho_2 S_{Y(2)} S_{X(2)}. \tag{91}$$

From equations (87) and (91), we have

$$E(\xi_Y^* \xi_X^*) = \lambda \rho S_Y S_X + \theta \rho_2 S_{Y(2)} S_{X(2)}. \tag{92}$$

Similarly, we can derive

$$E(\xi_Y^* \xi_X) = \lambda \rho S_Y S_X. \tag{93}$$

# Conditional density function for surrogate scalar response

Mounir Boumahdi <sup>1</sup>, Idir Ouassou <sup>2</sup>, Mustapha Rachdi <sup>3</sup>

## Abstract

This paper presents the estimator of the conditional density function of surrogated scalar response variable given a functional random one. We construct a conditional density function by using the available (true) response data and the surrogate data. Then, we build up some asymptotic properties of the constructed estimator in terms of the almost complete convergences. As a result, we compare our estimator with the classical estimator through the Relatif Mean Square Errors (RMSE). Finally, we end this analysis by displaying the superiority of our estimator in terms of prediction when we are lacking complete data.

**Key words:** Density function, surrogate response, functional variable, almost complete convergence, kernel estimators, scalar response, entropy, semi-metric space.

## 1. Introduction

There are many situations that may study the link between two variables, with the main goal to be able to predict new values. This predicted problem has been widely studied in the literature when both variables are of finite dimensions. Of course, the same problem can occur when some of the variables are functional. Our aim is to investigate this problem when the explanatory variable is functional and the response one is still real. We are based in the following model:

$$Y = m(X) + \varepsilon. \tag{1}$$

Where  $m$  is the regression operator,  $X$  is a functional covariate which belongs to a semi-metric space  $(E, d)$  and  $Y$  is the response variable,  $\varepsilon$  is a random error.

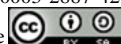
Our goal is to build the conditional density function for surrogate data by using the true response data and the surrogate data. By following the work of Wang (2006), Firas et al. (2019) and based on the work of Ramsay and Silverman (2002), Ferraty and Vieu (2006), Horvath and Kokoszka (2012), Cuevas (2014), Zhang (2014), Bongiorno et al. (2014), Hsing and Eubank (2015), Goia, Vieu (2016) and Wang, Chiou, and Müller (2016) and the references therein, we construct our estimator  $\hat{f}_R^X(y)$ .

The problem we are addressing in this work i.e., the unavailability of some data in the response variable, can be motivated both from a practical and a theoretical point of view.

<sup>1</sup>University Cadi Ayyad, The Eminent University of Science and Knowledge, Marrakech, Morocco. E-mail:mounirboumahdi1210112999@gmail.com. ORCID:<https://orcid.org/0000-0002-1017-3760>.

<sup>2</sup>University Cadi Ayyad, The Eminent University of Science and Knowledge, Marrakech, Morocco. E-mail:i.ouassou@uca.ac.ma. ORCID:<https://orcid.org/0000-0001-5573-5053>.

<sup>3</sup>Université Grenoble Alpes, Grenoble Cedex 09, France. E-mail: MUSTAPHA.RACHDI@univ-grenoble-alpes.fr. ORCID:<https://orcid.org/0000-0003-2887-4212>.



In fact, it may be difficult or expensive to exactly measure some response observations  $Y$ . Our goal is then to improve the modeling by filling/recovering some of the information missed in the response variable with this surrogate variable. In this case, one solution is to use the help of validation data to capture the underlying relation between the true variables and surrogate ones. Some examples where validation data are available can be found in Duncan and Hill (1985), Carroll and Wand (1991) and Pepe (1992).

This paper aims to study the conditional density for missing response by the kernel method, we explore in this work the aspect of missing data in the response variable to estimate the conditional density function for surrogate data. We adopt an approach based on validation data ideas. In fact, the idea is to introduce the information contained in both the validation data and the surrogate data.

The unavailable observation of  $Y$  will be replaced by the estimator of  $\mathbb{E}(Y | X, \tilde{Y})$ , denoted by  $U(X_j, \tilde{Y}_j)$  for all  $j \in \tilde{V}$  that corresponds to the size of the missing data, where  $\tilde{Y}$  is surrogate variable of  $Y$ . To estimate  $\mathbb{E}(Y | X, \tilde{Y})$  we adopt an approach based on validation data and the brut data (the primary data), which includes surrogate data and the corresponding observations of the covariate  $X$ .

Inside the simulation study of section 4, the surrogate variable  $\tilde{Y}_i$  of  $Y_i$  was generated from  $\tilde{Y}_i = \rho Z_i + \varepsilon_i$ , where  $Z_i$  is the standard score of  $Y_i$  and  $\varepsilon_i \sim N(0, \sqrt{1 - \rho^2})$ , in such a way that the correlation coefficient between  $Y_i$  and  $\tilde{Y}_i$  is approximately equal to  $\rho$  which would not be controllable in practice, but we can clearly notice that the quality of our  $\hat{f}_R^X$  depends on the size  $n$  of the validation data and  $\rho$ . Specifically, our estimator greatly better as the value of  $n$  and  $\rho$  increases.

We already know the convergence almost complete of the classical kernel estimator  $\hat{f}_C^X(y)$  (Ferraty and Vieu (2006)) towards the real  $f_Y^X(y)$ . In fact, within the section 4, we calculated and represented graphically the conditional density function estimator for surrogate data and we conduct a computational study on a simulated data in order to show advantages of using  $\hat{f}_R^X(y)$  over  $\hat{f}_V^X(y)$ .

Effectively, we are in a position to give the alternative estimator of  $\hat{f}_C^X(y)$  (estimator of Ferraty and Vieu) when we are lacking complete data with the help of  $\tilde{Y}$  (the surrogate variable of  $Y$ ), so in reality the choice of  $\tilde{Y}$  is important to improve the quality of our estimator. In practice we can cite as an example two diseases ( $Y$  and  $\tilde{Y}$ ) presenting similar symptoms, more that there is a strong correlation between these two diseases, more our estimator is better. So, there exists a wide scope of applied scientific fields for which our approach could be of interest for examples Biometrics, Genetics or Environmetrics and this approach can be helpful for a lot of statistical models when we are lacking complete data.

The main objective of this paper is to estimate the conditional density function for surrogate data. Then, we present the almost complete convergence of our estimator  $\hat{f}_R^X(y)$  and we study its performance against  $\hat{f}_V^X(y)$  by computing the relative mean squared error (RMSE) using simulated data.

## 2. Estimation procedure

Let  $(X, Y) \in \mathcal{F} \times \mathbb{R}$  denote a random vector, where  $(\mathcal{F}, d)$  is a semi-metric space equipped with the semi-metric  $d$ . We are concerned with the estimation of the conditional density function for surrogate data. Therefore, let  $(X_1, Y_1), \dots, (X_N, Y_N)$  be a random sample consisting of independent and identically distributed (i.i.d) variable from the distribution of  $(X, Y)$ .

The regression function for surrogate data defined in [?] as follows

$$\hat{m}_R(x) = \sum_{i \in V} Y_i W_{1,n,i}(x) + \sum_{j \in \bar{V}} U(X_j, \tilde{Y}_j) W_{1,n,j}(x), \tag{2}$$

with

$$U(X_j, \tilde{Y}_j) = \sum_{i \in V} Y_i W_{2,n,i}(X_j, \tilde{Y}_j), \quad \forall j \in \bar{V}. \tag{3}$$

We can estimate the conditional c.d.f  $F_Y^x(y)$  for surrogate data as follows

$$\hat{F}_R^x(y) = \sum_{i \in V} H\left(\frac{y - Y_i}{g}\right) W_{1,n,i}(x) + \sum_{j \in \bar{V}} R(X_j, y, \tilde{Y}_j) W_{1,n,j}(x), \tag{4}$$

where

$$W_{1,n,i}(x) = \frac{K\left(\frac{d(X_i, x)}{h}\right)}{\sum_{l=1}^N K\left(\frac{d(X_l, x)}{h}\right)}, \tag{5}$$

and

$$R(X_j, y, \tilde{Y}_j) = \sum_{i \in V} H\left(\frac{y - Y_i}{g}\right) W_{2,n,i}(X_j, \tilde{Y}_j), \quad \forall j \in \bar{V}. \tag{6}$$

With

$$W_{2,n,i}(X_j, \tilde{Y}_j) = \frac{W\left(\frac{d(X_j, X_i)}{h}, \frac{\tilde{Y}_i - \tilde{Y}_j}{b}\right)}{\sum_{l \in V} W\left(\frac{d(X_j, X_l)}{h}, \frac{\tilde{Y}_l - \tilde{Y}_j}{b}\right)}. \tag{7}$$

The conditional density function can be obtained by derivating the conditional c.d.f. Since we have now at hand some estimator  $\hat{F}_R^x(y)$  of  $F_Y^x(y)$ , it is natural to propose the following estimate:

$$\hat{f}_R^x(y) = \frac{\partial \hat{F}_R^x(y)}{\partial y}.$$

Assuming the differentiability of  $H$ , we build our new estimator of conditional density function for surrogate data as following:

$$\hat{f}_R^x(y) = \sum_{i \in V} \Omega_i(y) W_{1,n,i}(x) + \sum_{j \in \bar{V}} L(X_j, y, \tilde{Y}_j) W_{1,n,j}(x).$$

Where

$$\Omega_i(y) = g^{-1} K_0(g^{-1}(y - Y_i)),$$

and

$$L(X_j, y, \tilde{Y}_j) = \sum_{i \in V} g^{-1} K_0(g^{-1}(y - Y_i)) W_{2,n,i}(X_j, \tilde{Y}_j), \quad \forall j \in \bar{V}. \tag{8}$$

Where  $K$  is a kernel function and both  $h = h_N$  and  $g = g_N$  are a sequence of positive reals that tends to zero when  $N$  goes to infinity.

$$\forall u \in \mathbb{R}, \quad H(u) = \int_{-\infty}^u K_0(v) dv. \tag{9}$$

$K_0$  is a function from  $\mathbb{R}$  into  $\mathbb{R}^+$  such that  $\int K_0 = 1$ . To give an estimator of  $F_Y^x$  when there are surrogate data in the response variable, let us introduce the integer  $n$  ( $n < N$ ) that corresponds to the size of the validation set  $V$ . Let  $\bar{V}$  be the complementary set of  $V$  in the set  $\{1, 2, \dots, N\}$ .

$W$  is a kernel function which is defined on  $\mathbb{R}^2$  and  $b$  is sequence of real numbers which tend to zero. To simplify, we will use only one kernel. In sense that  $K = K_0$  and  $W(\cdot, \cdot) = K(\cdot)K(\cdot)$ . This consideration is because the choice of the kernel has less influence on the performance of the estimator.

### 3. Some asymptotic properties

In the sequel, when no confusion is possible, we will denote by  $C$  and  $C'$  some strictly positive generic constants, we denote by  $f^{x_1, \tilde{y}_1}(y)$  the conditional distribution function of  $Y$  given  $(X, \tilde{Y})$ :

$$f^{x_1, \tilde{y}_1}(y) = \frac{\partial F^{x_1, \tilde{y}_1}(y)}{\partial y},$$

with

$$F^{x_1, \tilde{y}_1}(y) = P(Y \leq y \mid x_1, \tilde{y}_1).$$

Recall that a semi-metric (sometimes called pseudo-metric) is just a metric violating the property  $[d(x, y) = 0] \Rightarrow [x = y]$ . We define the Kolmogorov's entropy as follows:

**Definition 3.1** Let  $S_{\mathcal{F}}$  be a subset of a semi-metric space  $\mathcal{F}$ , and let  $\varepsilon > 0$  be given. A finite set of point  $x_1, x_2, \dots, x_{n_0}$  in  $\mathcal{F}$  is called an  $\varepsilon$ -net for  $S_{\mathcal{F}}$  if  $S_{\mathcal{F}} \subset \bigcup_{k=1}^{n_0} B(x_k, \varepsilon)$ . The quantity  $\Psi_{S_{\mathcal{F}}} = \log(N_\varepsilon)$ , where  $N_\varepsilon$  is the minimal number of open balls in  $\mathcal{F}$  of radius  $\varepsilon$  which is necessary to cover  $S$ , is called the Kolmogorov's  $\varepsilon$ -entropy of the set  $S_{\mathcal{F}}$ .

This concept was introduced by Kolmogorov (see, Kolmogorov and Tikhomirov, 1959) and it represents a measure of the complexity of a set, in sense that, high entropy means that much information is needed to describe an element with an accuracy  $\varepsilon$ . Therefore, the choice of the topological structure (with other words, the choice of the semi-metric) will play a crucial role when one is looking the uniform (over  $S$ ) asymptotic results. For more precision about this concept, see Ferraty et al. (2010).

We consider the following assumptions:

(H1) For all  $x$  in the subset  $S_{\mathcal{F}}$  we have,

$$0 < C\phi(h) \leq P(X \in B(x, h)) \leq C'\phi(h) < \infty.$$



For all  $\tilde{y}$  in the subset  $S_{\mathcal{D}}$

$$0 < C\phi(b) \leq P(Y \leq \tilde{y} \leq Y + b) \leq C'\phi(b) < \infty.$$

For all  $x, \tilde{y}$  in the subset  $S_{\mathcal{F}} \times S_{\mathcal{D}}$

$$C\phi(h)\phi(b) < \mathbb{E}[K(h^{-1}d(x, X_i))K(b^{-1}(\tilde{y} - Y_1))] < C'\phi(h)\phi(b).$$

(H2) There exists  $b_1, b_2, b_3 > 0$  such that  $\forall x_1, x_2 \in S_{\mathcal{F}}, \forall y_1, y_2 \in S_{\mathcal{D}}$  and  $\forall \tilde{y}_1, \tilde{y}_2 \in S_{\mathcal{D}}$

$$|f^{x_1}(y_1) - f^{x_2}(y_1)| \leq C \left( d^{\beta_1}(x_1, x_2) + |y_1 - y_2|^{\beta_2} \right),$$

$$\text{and } |f^{x_1, \tilde{y}_1}(y_1) - f^{x_2, \tilde{y}_2}(y_1)| \leq C \left( d^{\beta_1}(x_1, x_2) + |y_1 - y_2|^{\beta_2} + |\tilde{y}_1 - \tilde{y}_2|^{\beta_3} \right).$$

(H3)  $K$  and  $K_0$  are bounded and Lipschitz kernel on its support  $[0, 1]$ , such that  $-\infty < C < K'(t) < C' < 0$ .

(H4) The functions  $\phi$  and  $\psi_{S_{\mathcal{F}}}$  are such that:

(H4a)  $\exists C > 0, \exists \eta_0 > 0, \forall \eta < \eta_0, \phi'(\eta) < C$ , and

$$\exists C > 0, \exists \eta_0 > 0, \forall 0 < \eta < \eta_0, \int_0^\eta \phi(u) du > C \eta \phi(\eta),$$

(H4b) For some  $\gamma \in (0, 1), \gamma' \in (0, 1)$  and  $\gamma'' \in (0, 1)$

$\lim_{n \rightarrow +\infty} n^\gamma h = \infty, \lim_{n \rightarrow +\infty} n^{\gamma'} g = \infty$  and  $\lim_{n \rightarrow +\infty} n^{\gamma''} b = \infty$ , and for  $n$  large enough:

$$\frac{(\log n)^2}{ng\phi(h)} < \frac{(\log n)^2}{ng\phi(b)\phi(h)} < \psi_{S_{\mathcal{F}}}\left(\frac{\log n}{n}\right) < \frac{ng\phi(b)\phi(h)}{\log n} < \frac{ng\phi(h)}{\log n}.$$

(H5) The Kolmogorov's  $\varepsilon$ -entropy of  $S_{\mathcal{F}}$  satisfies

$$\sum_{n=1}^\infty n^{2\gamma+1} \exp\left\{ (1-\beta)\psi_{S_{\mathcal{F}}}\left(\frac{\log n}{n}\right) \right\} < \infty, \text{ for some } \beta > 1,$$

and

$$\sum_{n=1}^\infty n^{2\gamma''+1} \exp\left\{ (1-\beta)\psi_{S_{\mathcal{F}}}\left(\frac{\log n}{n}\right) \right\} < \infty, \text{ for some } \beta > 1.$$

Note that (H4a) implies that for  $n$  large enough

$$0 \leq \phi(h) \leq Ch. \tag{10}$$

The condition (H4b) implies that:

$$\frac{\psi_{S_{\mathcal{F}}}(\varepsilon)}{ng\phi(h)} \rightarrow 0, \text{ and } \frac{\psi_{S_{\mathcal{F}}}(\varepsilon)}{ng\phi(b)\phi(h)} \rightarrow 0. \tag{11}$$

The condition (H4b) implies that:

$$\sum_{n=1}^{\infty} n^{2\gamma''+1} N_{\varepsilon}(\mathcal{S}_{\mathcal{F}})^{1-\beta} < \infty, \text{ and } \sum_{n=1}^{\infty} n^{2\gamma+1} N_{\varepsilon}(\mathcal{S}_{\mathcal{F}})^{1-\beta} < \infty. \tag{12}$$

Conditions (H2)-(H3) are very standard in the nonparametric setting. Concerning (H4a), the boundness of the derivative of  $\phi$  around zero allows to consider  $\phi$  as a Lipschitzian function. Hypothesis (H4b) deals with topological considerations by controlling the entropy of  $\mathcal{S}_{\mathcal{F}}$ . For a radius not too large, one requires that  $\psi_{\mathcal{S}_{\mathcal{F}}}\left(\frac{\log n}{n}\right)$  is not too small and not too large. Moreover (H4b) implies that  $\frac{\psi_{\mathcal{S}_{\mathcal{F}}}(\varepsilon)}{ng\phi(h)} \rightarrow 0$  and  $\frac{\psi_{\mathcal{S}_{\mathcal{F}}}(\varepsilon)}{ng\phi(h)\phi(b)} \rightarrow 0$  tends to 0 when  $n$  tends to  $+\infty$ , in some ‘‘usual’’ cases, one has  $\psi_{\mathcal{S}_{\mathcal{F}}}\left(\frac{\log n}{n}\right) \sim C \log n$ . The assumption (H5) acts on the Kolmogorov  $\varepsilon$ -entropy of  $\mathcal{S}_{\mathcal{F}}$ .

The following Theorem states the rate of convergence of  $\hat{f}_R^x$  for the surrogated scalar response, uniformly over the set  $\mathcal{S}_{\mathcal{F}}$  and  $\mathcal{S}_{\mathcal{R}}$ . The asymptotics are stated in terms of almost complete convergence (denoted by a.co.), which imply both weak and strong convergences (see Section A-1 in Ferraty and Vieu, 2006).

**Theorem 3.1** *Under the hypotheses (H1)-(H5), we have*

$$\begin{aligned} \sup_{x \in \mathcal{S}_{\mathcal{F}}} \sup_{y \in \mathcal{S}_{\mathcal{R}}} |\hat{f}_R^x(y) - f_Y^x(y)| &= O(h^{\beta_1}) + O(g^{\beta_2}) \\ &+ O_{a.co.} \left( \sqrt{\frac{\psi_{\mathcal{S}_{\mathcal{F}}}\left(\frac{\log n}{n}\right)}{ng\phi(h)}} \right) + O_{a.co.} \left( \sqrt{\frac{\psi_{\mathcal{S}_{\mathcal{F}}}\left(\frac{\log N}{N}\right)}{Ng\phi(h)}} \right) \\ &+ O_{a.co.} \left( \sqrt{\frac{\psi_{\mathcal{S}_{\mathcal{F}}}\left(\frac{\log n}{n}\right)}{n\phi(h)\phi(b)}} \right) + O_{a.co.} \left( \sqrt{\frac{\psi_{\mathcal{S}_{\mathcal{F}}}\left(\frac{\log n}{n}\right)}{ng\phi(h)\phi(b)}} \right). \end{aligned}$$

### 4. Numerical results

In this section, we evaluate the performance of the proposed estimator by conducting a number of simulation studies. Let  $\hat{f}_V^x(y)$  be the classical conditional density function estimator which is obtained with the true observations in the validation data set  $V$ :

$$\hat{f}_V^x(y) = \frac{\sum_{i \in V} K(h^{-1}d(x, X_i))g^{-1}K_0(g^{-1}(y - Y_i))}{\sum_{i \in V} K(h^{-1}d(x, X_i))},$$

and  $\hat{f}_C^x(y)$  the classical kernel estimator which is obtained with the complete data for (such as an example with  $N = 300$  in the simulation below)

$$\hat{f}_C^x(y) = \frac{\sum_{i=1}^N K(h^{-1}d(x, X_i))g^{-1}K_0(g^{-1}(y - Y_i))}{\sum_{i=1}^N K(h^{-1}d(x, X_i))}.$$

Within this section we will calculate and represent graphically the conditional density function estimator for surrogate data and we conduct a computational study on a simulated data in order to show advantages of using  $\hat{f}_R^x(y)$  over  $\hat{f}_V^x(y)$ .

We choose  $K$  and  $K_0$  the Gaussian kernel as follows

$$K_0(u) = K(u) = \frac{1}{\sqrt{2\pi}} \exp^{-u^2/2}.$$

We generate 400 observations  $(X_i, Y_i)_i$  using the following model:

$$Y_i = m(X_i) + \varepsilon.$$

Where the errors  $\varepsilon_i$  are i.i.d. according to the normal distribution  $N(0;5)$ . More precisely, the functional regressors  $X_i(t)$  are defined, for any  $t \in [0, 1]$ , by

$$X_i(t) = \sin(2\pi t) + W_i * t.$$

Where  $W_i \sim U(0.5;2)$ . The scalar response variable  $Y$  is generated by taking as a regression operator:

$$m(x) = 2\pi * \sin(b_i) \times \int_0^1 x^2(t)dt + \varepsilon.$$

Where:  $\varepsilon_i \sim N(0, 2)$  and  $b_i \sim N(0, 0.1)$ . Let  $I_0 = \{1, \dots, 300\}$  and  $I_1 = \{301, \dots, 400\}$  be two subsets of indices. Then, we choose  $\Delta = (X_i, Y_i)_{i \in I_0}$  as the learning sample and  $\Gamma = \{(X_i, Y_i)\}_{i \in I_1}$  as the testing sample. The surrogate variable  $\tilde{Y}_i$  of  $Y_i$ , for all  $i \in I_0$  was generated from  $\tilde{Y}_i = \rho Z_i + \varepsilon_i$ , where  $Z_i$  is the standard score of  $Y_i$  and  $\varepsilon_i \sim N(0, \sqrt{1 - \rho^2})$ , in such a way that the correlation coefficient between  $Y_i$  and  $\tilde{Y}_i$  is approximately equal to  $\rho$  which would not be controllable in practice.

In the sequel of this simulation study, we take  $\rho = 0.75$ . From the learning sample containing  $N = 300$  functional data, we randomly choose a set  $V$  of  $n$  validation data  $\{(X_i, Y_i)\}_{i \in V}$  which allows to build the estimator  $\hat{f}_V^x(y)$  of  $f_Y^x(y)$ .

The estimator  $\hat{f}_R^x(y)$  is then constructed by using the surrogate data  $\{(X_i, Y_i)\}_{i \in \bar{V}}$  with the help of the validation data, where  $\bar{V} \cup V = \{1, \dots, N\}$ . It should be pointed out that for  $N = n$  (complete observations), we have  $\hat{f}_V^x(y) = \hat{f}_R^x(y) = \hat{f}_C^x(y)$ .

We evaluate the performance of the estimator  $\hat{f}_R^x(y)$  in terms of prediction, by computing the relative mean squared error (RMSE) on the test sample:

$$RMSE(\hat{f}_R^x) = \sqrt{\frac{\sum_{i \in \Gamma} (\hat{f}_C^x(Y_i) - \hat{f}_R^x(Y_i))^2}{100}}.$$

We have run 100 replicates of the simulation process for various values of  $n$ . We computed, for the two estimators  $\hat{f}_R^x(y)$  and  $\hat{f}_V^x(y)$  the mean and relative mean squared error (RMSE) over this 100 replications.

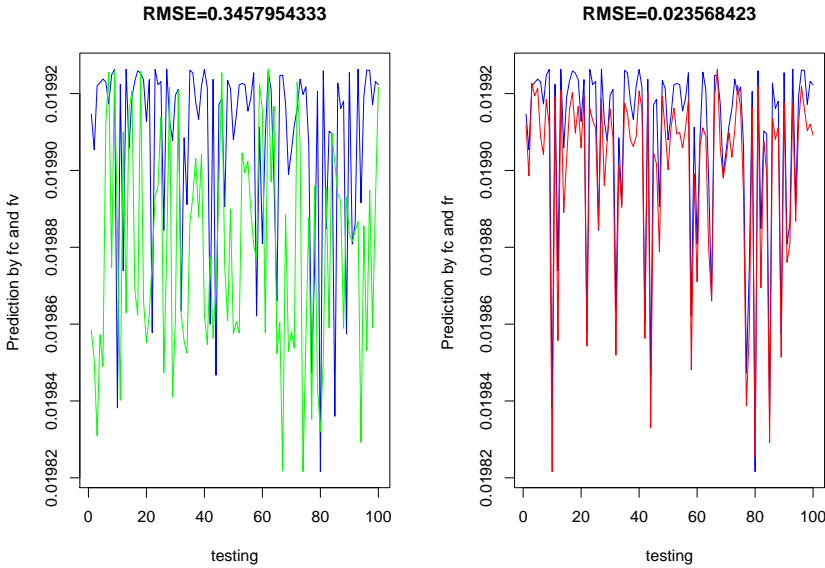
The comparison study results, for different values of percentage of validation data in samples:

$$p(V) = \frac{\text{card}(V)}{N} \times 100\% = \frac{n}{N} \times 100\%.$$

The results are summarized in the following Table 1.

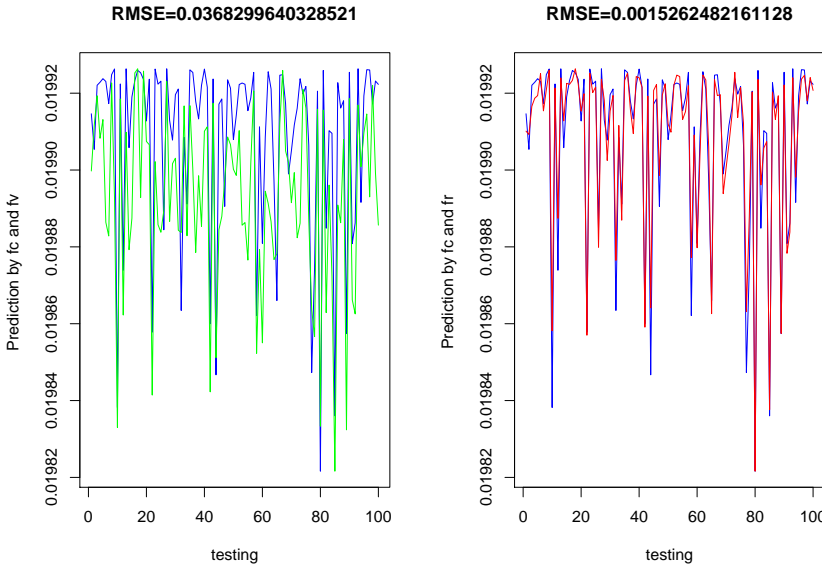
**Table 1.**  $\hat{f}_R^x(Y_i)$  and  $\hat{f}_V^x(Y_i)$  whereas  $\hat{f}_C^x(Y_i)$  for  $n = 100$  and  $n = 210$ .

estimator	$p(V)$	Mean	RMSE
$\hat{f}_V^x(y)$	33%	0.048	0.34
$\hat{f}_R^x(y)$	33%	0.03767	0.02356
$\hat{f}_C^x(y)$	-	0.03960	-
$\hat{f}_V^x(y)$	70%	0.044	0.0368
$\hat{f}_R^x(y)$	70%	0.03882	0.001
$\hat{f}_C^x(y)$	-	0.03960	-



**Figure 1.**  $\hat{f}_R^x$  (the red line ) and  $\hat{f}_V^x$  (the green line ) with  $\hat{f}_C^x$  (blue line) for  $\text{Card}(V)=n=100$ .

Obviously the quality of the prediction of the two estimators depend on the size  $n$  of the validation data. Specifically, RMSE decrease when the value of  $n$  increases. On the other hand, for  $n = 100$  that means the percentage of validation data in a sample is 33% our estimator  $\hat{f}_R^x(y)$  is better than  $\hat{f}_V^x(y)$  in terms of RMSE inferior. In addition for  $n = 210$  that means that we know 70% of data, our  $\hat{f}_R^x(y)$  still greatly better as result of  $RMSE = 0.001$ . Nearly with the same mean of  $\hat{f}_C^x(y)$ .



**Figure 2.**  $\hat{f}_R^x$  (the red line) and  $\hat{f}_V^x$  (the green line) with  $\hat{f}_C^x$  (blue line) for  $Card(V) = n = 210$ .

It can be noticed from Figure 1 and Figure 2 that our  $\hat{f}_R^x(y)$  is closer than  $\hat{f}_V^x(y)$  to the curve  $\hat{f}_C^x(y)$  which represents the estimator with the complete samples. Consequently, even if the percentage of validation data in sample increases from 33% to 70%, the estimator  $\hat{f}_R^x(y)$  keeps performing better than  $\hat{f}_V^x$ .

### 5. Remarks and Conclusion

This paper has stated uniform consistency results when  $X$  is functional and  $Y$  is scalar. The fact to be able to state results on the quantity

$$\sup_{x \in S_{\mathcal{F}}} \sup_{y \in S_{\mathcal{R}}} |\hat{f}_R^x(y) - f_Y^x(y)|.$$

Allows directly to obtain results on quantity

$$|\hat{f}_R^x(y) - f_Y^x(y)|.$$

The entropy function represents a measure of the complexity of a set, in sense that, high entropy means that much information is needed to describe an element with an accuracy  $\epsilon = \frac{\log n}{n}$ , in fact, the quality of the prediction of this estimator depends on the size  $n$  of the validation data. By building a suitable projection-based semi-metric, the entropy function becomes  $\psi_{S_{\mathcal{F}}} \left( \frac{\log n}{n} \right) = O(\log n)$  and for  $N = n$  (without surrogate data) we get the estimator of Ferraty and Vieu (2006)

$$\sup_{x \in S_{\mathcal{F}}} \sup_{y \in S_{\mathcal{R}}} |\hat{f}_R^x(y) - f_Y^x(y)| = O(h^{\beta_1}) + O(g^{\beta_2}) + O_{a.co.} \left( \sqrt{\frac{\log n}{ng\phi(h)}} \right).$$

We present in this paper the almost complete convergence of conditional density function for surrogated scalar response variable given a functional random by using validation sample set. In addition, we show the performance of our estimator  $\hat{f}_R^x(y)$  than  $\hat{f}_V^x(y)$  to reduce RMSE by using the simulated data. This confirms that our estimator is a good alternative to the  $\hat{f}_C^x(y)$  estimator (see Ferraty and Vieu, 2006) when we lack complete data.

**Proof of Theorem**

We note that :

$$i \in V \Rightarrow i \in \{1, \dots, n\}, \text{ and } j \in \bar{V} \Rightarrow j \in \{n + 1, \dots, N\}.$$

We can write

$$\begin{aligned} \hat{f}_R^x(y) - f_Y^x(y) &= \sum_{i \in V} \Omega_i(y) W_{1,n,i}(x) - \sum_{i \in V} f_Y^{X_i, \tilde{Y}_i}(y) W_{1,n,i}(x) \\ &\quad - \sum_{j \in \bar{V}} f_Y^{X_j, \tilde{Y}_j}(y) W_{1,n,j}(x) + \sum_{j \in \bar{V}} L(X_j, \tilde{Y}_j) W_{1,n,j}(x) \\ &\quad + \sum_{i=1}^N f_Y^{X_i, \tilde{Y}_i}(y) W_{1,n,i}(x) - f_Y^x(y). \\ &= E_1 + E_2 + E_3, \end{aligned}$$

with

$$\begin{cases} E_1 &= \sum_{i \in V} \left( \Omega_i(y) - f_Y^{X_i, \tilde{Y}_i}(y) \right) W_{1,n,i}(x), \\ E_2 &= \sum_{j \in \bar{V}} \left( L(X_j, \tilde{Y}_j) - f_Y^{X_j, \tilde{Y}_j}(y) \right) W_{1,n,j}(x), \\ E_3 &= \sum_{i=1}^N \left( f_Y^{X_i, \tilde{Y}_i}(y) - f_Y^x(y) \right) W_{1,n,i}(x). \end{cases}$$

And

$$\Omega_i(y) = g^{-1} K_0(g^{-1}(y - Y_i)).$$

Furthermore, we put

$$\Delta_i(x) = \frac{K\left(\frac{d(X_i, x)}{h}\right)}{\mathbb{E}\left[K\left(\frac{d(X_i, x)}{h}\right)\right]},$$

and we define

$$\begin{cases} \hat{r}_1(x) &= \frac{1}{n} \sum_{i \in V} \Delta_i(x), \\ \tilde{r}_1(x) &= \frac{1}{N} \sum_{i=1}^N \Delta_i(x), \\ \hat{r}_2(x, y) &= \frac{1}{n} \sum_{i \in V} \left( \Omega_i(y) - f_Y^{X_i, \tilde{Y}_i}(y) \right) \Delta_i(x), \\ \hat{r}_3(x) &= \frac{1}{N} \sum_{i=1}^N \left( f_Y^{X_i, \tilde{Y}_i}(y) - f_Y^x(y) \right) \Delta_i(x). \end{cases}$$

By the definition of  $\hat{r}_1, \tilde{r}_1, \hat{r}_2$ , and  $\hat{r}_3$  we have:

$$E_1 = \frac{1}{\hat{r}_1(x)} (\hat{r}_2(x, y) - \mathbb{E}(\hat{r}_2(x, y))) + \frac{\mathbb{E}(\hat{r}_2(x, y))}{\hat{r}_1(x)},$$

and

$$E_3 = \frac{1}{\tilde{r}_1(x)} (\hat{r}_3(x, y) - \mathbb{E}(\hat{r}_3(x, y))) + \frac{\mathbb{E}(\hat{r}_3(x, y))}{\tilde{r}_1(x)}.$$

The numerators in this decomposition will be treated directly by using Lemma 6.2 and Lemma 6.3 below, while the denominators are treated directly by using Lemma 6.1 together with part i) of Proposition A.6 defined in p232 of Ferraty and Vieu, 2006. For the term  $E_2$  will be treated by using Lemma 6.4.

Finally, the Theorem 3.2 is consequence of the following intermediate results

**Lemma 5.1** *Under the hypotheses (H1) and (H3)-(H5), we have*

$$\sup_{x \in S_{\mathcal{F}}} |\hat{r}_1(x) - 1| = O_{a.co.} \left( \sqrt{\frac{\Psi_{S_{\mathcal{F}}} \left( \frac{\log n}{n} \right)}{n\phi(h)}} \right),$$

and

$$\sum_{n=1}^{\infty} P \left( \inf_{x \in S_{\mathcal{F}}} \hat{r}_1(x) < \frac{1}{2} \right) < \infty.$$

The Proof of this Lemma is detailed in [?]

**Lemma 5.2** *Under the hypotheses (H1),(H2) and (H4)-(H5), we have*

$$\sup_{x \in S_{\mathcal{F}}} \sup_{y \in S_{\mathcal{H}}} |\mathbb{E}[\hat{r}_2(x, y)]| = O \left( g^{\beta_2} \right),$$

and

$$\sup_{x \in S_{\mathcal{F}}} \sup_{y \in S_{\mathcal{H}}} |\mathbb{E}[\hat{r}_3(x, y)]| = O \left( h^{\beta_1} \right).$$

**Proof of Lemma 5.2**

By stationarity, we have

$$\begin{aligned} |\mathbb{E}[\hat{r}_2(x, y)]| &= \left| \mathbb{E} \left[ \Delta_1(x) \mathbb{E} \left[ \left( \Omega_1(y) - f_Y^{X_1, \tilde{Y}_1}(y) \right) | X_1 \right] \right] \right| \\ &= \left| \mathbb{E} \left[ \Delta_1(x) \mathbb{E} [\Omega_1(y) | X_1] - \mathbb{E} [f_Y^{X_1, \tilde{Y}_1}(y) | X_1] \right] \right| \\ &= \left| \mathbb{E} \left[ \mathbf{1}_{B(x, h)}(X_1) \Delta_1(x) \mathbb{E} [\Omega_1(y) | X_1] - f_Y^{X_1}(y) \right] \right|. \end{aligned}$$

The fact that  $\int_{\mathbb{R}} K_0(u) du = 1$  allows us to write:

$$\begin{aligned} \mathbb{E} [\Omega_1(y) | X_1] - f_Y^{X_1}(y) &= \int_{\mathbb{R}} g^{-1} K_0(g^{-1}(y - u)) \left( f_Y^{X_1}(y) - f_Y^{X_1}(u) \right) du \\ &= \int_{\mathbb{R}} K_0(v) \left( f_Y^{X_1}(y) - f_Y^{X_1}(y - vg) \right) dv \end{aligned}$$

Thus, under (H3) we obtain uniformly

$$\left| \mathbb{E}[\Omega_1(y)|X_1] - f_Y^{X_1}(y) \right| \leq Cg^{\beta_2}.$$

Hence, we get

$$\forall x \in S_{\mathcal{F}}, \quad |\mathbb{E}[\hat{r}_2(x, y)]| \leq Cg^{\beta_2}.$$

$$\begin{aligned} |\mathbb{E}[\hat{r}_3(x, y)]| &= \left| \mathbb{E} \left[ \Delta_1(x) \mathbb{E} \left[ \left( f_Y^{X_i, \tilde{Y}_i}(y) - f_Y^x(y) \right) | X_1 \right] \right] \right| \\ &= \mathbb{E} \left[ \mathbf{1}_{B(x, h)}(X_1) \Delta_1(x) \left| f_Y^{X_1}(y) - f_Y^x(y) \right| \right] \leq Ch^{\beta_1}. \end{aligned}$$

**Lemma 5.3** *Under the assumptions of the Theorem, we have*

$$\sup_{x \in S_{\mathcal{F}}} \sup_{y \in S_{\mathcal{D}}} |\hat{r}_2(x, y) - \mathbb{E}[\hat{r}_2(x, y)]| = O_{a.co.} \left( \sqrt{\frac{\Psi_{S_{\mathcal{F}}} \left( \frac{\log n}{n} \right)}{ng\phi(h)}} \right),$$

and

$$\sup_{x \in S_{\mathcal{F}}} \sup_{y \in S_{\mathcal{D}}} |\hat{r}_3(x, y) - \mathbb{E}[\hat{r}_3(x, y)]| = O_{a.co.} \left( \sqrt{\frac{\Psi_{S_{\mathcal{F}}} \left( \frac{\log N}{N} \right)}{Ng\phi(h)}} \right).$$

### Proof of Lemma 5.3

We treat only the first case, the second result can be treated by the same arguments. Firstly, we simplify the notation by denoting for all  $i = 1, \dots, n$ , by

$$K_i(x) = K(h^{-1}d(x, X_i)).$$

Observe that, according to (H1) and (H3) we have

$$\forall x \in S_{\mathcal{F}} \quad C\phi(h) < \mathbb{E}[K_1(x)] < C'\phi(h). \quad (13)$$

Next, we denote by  $x_1, \dots, x_{N_\varepsilon(S_{\mathcal{F}})}$  an  $\varepsilon$ -net (see Kolmogorov and Tikhomirov (1959)) for  $S_{\mathcal{F}}$  and by  $t_1, \dots, t_{d_n}$  some  $l_n$ -net for the compact  $S_{\mathcal{D}}$ . Furthermore, for all  $x$  in  $S_{\mathcal{F}}$  and  $y$  in  $S_{\mathcal{D}}$  we put

$$k(x) = \arg \min_{k \in \{1, 2, \dots, N_\varepsilon(S_{\mathcal{F}})\}} d(x, x_k) \text{ and } j(y) = \arg \min_{j \in \{1, 2, \dots, d_n\}} |y - t_j|.$$



Now, we fix  $\varepsilon = \frac{\log n}{n}$  and  $l_n = n^{-2\gamma-1}$  and we use the following decomposition

$$\begin{aligned} |\hat{r}_2(x,y) - \mathbb{E}[\hat{r}_2(x,y)]| &\leq \underbrace{\sup_{x \in S_{\mathcal{F}}} \sup_{y \in S_{\mathcal{R}}} |\hat{r}_2(x,y) - \hat{r}_2(x_{k(x)},y)|}_{T_1} \\ &+ \underbrace{\sup_{x \in S_{\mathcal{F}}} \sup_{y \in S_{\mathcal{R}}} |\hat{r}_2(x_{k(x)},y) - \hat{r}_2(x_{k(x)},t_{j(y)})|}_{T_2} \\ &+ \underbrace{\sup_{x \in S_{\mathcal{F}}} \sup_{y \in S_{\mathcal{R}}} |\hat{r}_2(x_{k(x)},t_{j(y)}) - \mathbb{E}[\hat{r}_2(x_{k(x)},t_{j(y)})]|}_{T_3} \\ &+ \underbrace{\sup_{x \in S_{\mathcal{F}}} \sup_{y \in S_{\mathcal{R}}} |\mathbb{E}[\hat{r}_2(x_{k(x)},t_{j(y)})] - \mathbb{E}[\hat{r}_2(x_{k(x)},y)]|}_{T_4} \\ &+ \underbrace{\sup_{x \in S_{\mathcal{F}}} \sup_{y \in S_{\mathcal{R}}} |\mathbb{E}[\hat{r}_2(x_{k(x)},y)] - \mathbb{E}[\hat{r}_2(x,y)]|}_{T_5}. \end{aligned}$$

For the term  $T_1$  we employ the Lipschitzianity of the kernel  $K$  on  $[0, 1]$  which gives

$$T_1 \leq \frac{C}{n} \sum_{i=1}^n Z_i \text{ with } Z_i = \frac{\varepsilon}{hg\phi(h)} \mathbb{1}_{B(x,h) \cup B(x_{k(x)},h)}(X_i),$$

Therefore, it is clear that the assumption (H3) permits to write that

$$Z_1 = O\left(\frac{\varepsilon}{h\phi(h)}\right), \mathbb{E}[Z_1] = O\left(\frac{\varepsilon}{hg}\right) \text{ and } \text{var}(Z_1) = O\left(\frac{\varepsilon^2}{h^2g^2\phi(h)}\right).$$

So, we get

$$\mathbb{E}(|Z_1|^m) \leq \frac{C\varepsilon^m}{h^m g^m \phi(h)^{m-1}}. \tag{14}$$

By using the result (10) together with the definition of  $\varepsilon$  we have for  $n$  large enough:

$$\frac{\varepsilon}{hg} \leq C.$$

So, we get:

$$\mathbb{E}(|Z_1|^m) \leq \frac{C\varepsilon^{m-1}}{h^{m-1}g^{m-1}\phi(h)^{m-1}}.$$

Now, by applying Corollary A.8 in Ferraty and Vieu (2006) with  $a^2 = \frac{\varepsilon}{hg\phi(h)}$ , we get:

$$\frac{1}{n} \sum_{i=1}^n Z_i = EZ_1 + O_{a.co.} \left( \sqrt{\frac{\varepsilon \log n}{ngh\phi(h)}} \right).$$

Finally, applying (14) for  $m = 1$  one gets

$$T_1 = O\left(\frac{\varepsilon}{hg}\right) + O_{a.co.} \left( \sqrt{\frac{\varepsilon \log n}{nhg\phi(h)}} \right).$$

By (10) and the definition of  $\varepsilon$  for  $n$  large enough :

$$C \frac{(\log n)^2}{(ng\phi(h))^2} \geq \frac{\varepsilon \log n}{nhg\phi(h)}$$

Using (H4b) together with (11) and the fact that:

$$\left\{ \left| \frac{1}{n} \sum_{i=1}^n Z_i \right| > \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{ng\phi(h)}} \right\} \subset \left\{ \left| \frac{1}{n} \sum_{i=1}^n Z_i \right| > \sqrt{\frac{C(\log n)^2}{(ng\phi(h))^2}} \right\},$$

we get

$$T_1 = O_{a.co.} \left( \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{ng\phi(h)}} \right). \quad (15)$$

Thus, by Assumption (H4b) we deduce that

$$T_1 = O_{a.co.} \left( \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{ng\phi(h)}} \right) \text{ and } T_5 = O \left( \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{ng\phi(h)}} \right). \quad (16)$$

We use the same ideas to treat  $R_2$ . In fact, we use the Lipschitz condition on the kernel  $K$  and the assumption (H2) to write that

$$\begin{aligned} |\hat{r}_2(x_{k(x)}, y) - \hat{r}_2(x_{k(x)}, t_{j(y)})| &\leq \frac{C}{n\phi(h)} \sum_{i=1}^n K_i(x_{k(x)}) (|\Omega_i(y) - \Omega_i(t_{j(y)})| \\ &\quad + |f_Y^{X_i, \tilde{Y}_i}(y) - f_Y^{X_i, \tilde{Y}_i}(t_{j(y)})|) \\ &\leq \frac{C}{n} \sum_{i=1}^n Z_i, \end{aligned}$$

where  $Z_i = \frac{l_n K_i(x_{k(x)}) \mathbb{I}_{B(x_{k(x)}, h)}(X_i)}{g^2 \phi(h)}$ .

It is clear that the assumption (H3) permits to write that

$$Z_1 = O \left( \frac{l_n}{g^2 \phi(h)} \right), \mathbb{E}[Z_1] = O \left( \frac{l_n}{g^2} \right) \text{ and } \text{var}(Z_1) = O \left( \frac{l_n^2}{g^4 \phi(h)} \right).$$

Invoking the same idea in (15), allows to get:

$$T_2 = O_{a.co.} \left( \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{gn\phi(h)}} \right) \text{ and } T_4 = O \left( \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{gn\phi(h)}} \right). \quad (17)$$

It remains to evaluate  $R_3$ . Indeed, we write

$$\begin{aligned} P\left(T_3 > \eta \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{n \phi(h)}}\right) &= P\left(\max_{j \in \{1, 2, \dots, d_n\}} \max_{k \in \{1, \dots, N_{\varepsilon}(S_{\mathcal{F}})\}} |\hat{r}_2(x_k, t_j) - \mathbb{E}\hat{r}_2(x_k, t_j)| > \eta \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{n \phi(h)}}\right) \\ &\leq d_n N_{\varepsilon}(S_{\mathcal{F}}) \max_{j \in \{1, 2, \dots, d_n\}} \max_{k \in \{1, \dots, N_{\varepsilon}(S_{\mathcal{F}})\}} P\left(|\hat{r}_2(x_k, t_j) - \mathbb{E}\hat{r}_2(x_k, t_j)| > \eta \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{n \phi(h)}}\right) \\ &\leq d_n N_{\varepsilon}(S_{\mathcal{F}}) \max_{j \in \{1, 2, \dots, d_n\}} \max_{k \in \{1, \dots, N_{\varepsilon}(S_{\mathcal{F}})\}} P\left(\left|\frac{1}{n} \sum_{i=1}^n \Gamma_i\right| > \eta \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{n \phi(h)}}\right). \end{aligned}$$

Where

$$\Gamma_i = \frac{1}{\mathbb{E}[K_1(x)]} \left[ K_i(x_k)(\Omega_i(t_j) - f^{X_i, \tilde{Y}_i}(t_j)) - E\left(K_i(x_k)(\Omega_i(t_j) - f^{X_i, \tilde{Y}_i}(t_j))\right) \right].$$

It follows, from the fact that the kernel  $K$  and  $K_0$  and  $f^{X_i, \tilde{Y}_i}$  are bounded, that

$$E|\Gamma_i|^2 \leq C(\phi(h))^{-1}.$$

Thus, we apply the Bernstein exponential inequality, we obtain for all  $j \leq d_n$ , that

$$P\left(|\hat{r}_2(x_k, t_j) - \mathbb{E}\hat{r}_2(x_k, t_j)| > \eta \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{n \phi(h)}}\right) \leq 2 \exp\left\{-C\eta^2 \Psi_{S_{\mathcal{F}}}(\varepsilon)\right\}.$$

Therefore, by choosing  $C\eta^2 = \beta$ , and using the fact that  $d_n = O(l_n^{-1})$ , we conclude that

$$\begin{aligned} d_n N_{\varepsilon}(S_{\mathcal{F}}) \max_{j \in \{1, 2, \dots, d_n\}} \max_{k \in \{1, \dots, N_{\varepsilon}(S_{\mathcal{F}})\}} P\left(|\hat{r}_2(x_k, t_j) - \mathbb{E}\hat{r}_2(x_k, t_j)| > \eta \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{n \phi(h)}}\right) \\ \leq C' d_n (N_{\varepsilon}(S_{\mathcal{F}}))^{1-C\eta^2}. \end{aligned}$$

Finally, we obtain

$$T_3 = O_{a.co.} \left( \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{n^{1-\gamma} \phi(h)}} \right) \tag{18}$$

For the term  $\hat{r}_3(x, y) - \mathbb{E}[\hat{r}_3(x, y)]$ . First we fix  $\varepsilon = \frac{\log N}{N}$  and  $l_N = N^{-2\gamma-1}$ .

Using the decomposition and invoking the same arguments as for the proof of  $\hat{r}_2(x, y) - \mathbb{E}[\hat{r}_2(x, y)]$ , we get:

$$\sup_{x \in S_{\mathcal{F}}} \sup_{y \in S_{\mathcal{F}}} |\hat{r}_3(x, y) - \mathbb{E}[\hat{r}_3(x, y)]| = O_{a.co.} \left( \sqrt{\frac{\Psi_{S_{\mathcal{F}}}\left(\frac{\log N}{N}\right)}{Ng\phi(h)}} \right).$$

**Lemma 5.4** Under the assumptions of Theorem (H1)-(H6), we have  $\forall j \in \bar{V}$

$$\begin{aligned} \sup_{x \in S_{\mathcal{F}}} \sup_{y \in S_{\mathcal{F}}} |f^{\tilde{Y}_j, X_j}(y) - L(X_j, \tilde{Y}_j)| &= O(h^{\beta_1}) + O_{a.co.} \left( \sqrt{\frac{\psi_{S_{\mathcal{F}}} \left( \frac{\log n}{n} \right)}{n\phi(h)\phi(b)}} \right) \\ &+ O_{a.co.} \left( \sqrt{\frac{\psi_{S_{\mathcal{F}}} \left( \frac{\log n}{n} \right)}{ng\phi(b)\phi(h)}} \right). \end{aligned}$$

To simplify we put  $X_j = x$  and  $\tilde{Y}_j = \tilde{y}$ .

**Proof of Lemma 5.4**

The proof is based on the following decomposition

$$\begin{aligned} L(x, y, \tilde{y}) - f_Y^{X_i, \tilde{y}_i}(y) &= \frac{1}{L_1(x, \tilde{y})} \left[ L_2(x, y, \tilde{y}) - \mathbb{E}[L_2(x, y, \tilde{y})] \right] \\ &+ \frac{1}{L_1(x, \tilde{y})} \left[ \mathbb{E}[L_2(x, y, \tilde{y})] - f_Y^{X_i, \tilde{y}_i}(y) \right] + [1 - L_1(x, \tilde{y})] \frac{f_Y^{X_i, \tilde{y}_i}(y)}{L_1(x, \tilde{y})}. \end{aligned}$$

Where

$$L_1(x, \tilde{y}) = \frac{1}{n\mathbb{E}[K(h^{-1}d(x, X_1))K(b^{-1}(\tilde{y} - \tilde{Y}_1))]} \sum_{i \in \bar{V}} K(h^{-1}d(x, X_i))K(b^{-1}(\tilde{y} - \tilde{Y}_i)),$$

and

$$L_2(x, y, \tilde{y}) = \frac{1}{n\mathbb{E}[K(h^{-1}d(x, X_1))K(b^{-1}(\tilde{y} - \tilde{Y}_1))]} \sum_{i \in \bar{V}} K(h^{-1}d(x, X_i))K(b^{-1}(\tilde{y} - \tilde{Y}_i))\Omega_i(y).$$

$$\begin{aligned} |L_1(x, \tilde{y}) - \mathbb{E}[L_1(x, \tilde{y})]| &\leq \underbrace{\sup_{x \in S_{\mathcal{F}}} \sup_{\tilde{y} \in S_{\mathcal{D}}} |L_1(x, \tilde{y}) - L_1(x_k, \tilde{y})|}_{R_1} \\ &+ \underbrace{\sup_{x \in S_{\mathcal{F}}} \sup_{\tilde{y} \in S_{\mathcal{D}}} |L_1(x_k, \tilde{y}) - L_1(x_k, t_{j(\tilde{y})})|}_{R_2} \\ &+ \underbrace{\sup_{x \in S_{\mathcal{F}}} \sup_{\tilde{y} \in S_{\mathcal{D}}} |L_1(x_k, t_{j(\tilde{y})}) - \mathbb{E}[L_1(x_k, t_{j(\tilde{y})})]|}_{R_3} \\ &+ \underbrace{\sup_{x \in S_{\mathcal{F}}} \sup_{\tilde{y} \in S_{\mathcal{D}}} |\mathbb{E}[L_1(x_k, t_{j(\tilde{y})})] - \mathbb{E}[L_1(x_k, \tilde{y})]|}_{R_4} \\ &+ \underbrace{\sup_{x \in S_{\mathcal{F}}} \sup_{\tilde{y} \in S_{\mathcal{D}}} |\mathbb{E}[L_1(x_k, \tilde{y})] - \mathbb{E}[L_1(x, \tilde{y})]|}_{R_5}. \end{aligned}$$

For the term  $R_1$  we employ the Lipschitzianity of the kernel  $K$  on  $[0, 1]$  with (H1) and (H2) lead directly

$$R_1 \leq \frac{C}{n} \sum_{i=1}^n Z_i \text{ with } Z_i = \frac{\varepsilon}{h\phi(h)\phi(b)} \mathbb{I}_{B(x,h) \cup B(x_k(x),h)}(X_i) \mathbb{I}_{\{Y \leq \bar{y} \leq Y+b\}},$$

It is clear that the assumption (H3) permits to write that

$$Z_1 = O\left(\frac{\varepsilon}{h\phi(h)}\right), \mathbb{E}[Z_1] = O\left(\frac{\varepsilon}{h}\right) \text{ and } \text{var}(Z_1) = O\left(\frac{\varepsilon^2}{h^2\phi(b)\phi(h)}\right).$$

By using the same steps as (15) we get

$$R_1 = O_{a.co.} \left( \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{n\phi(h)\phi(b)}} \right) \text{ and } R_5 = O \left( \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{n\phi(h)\phi(b)}} \right). \tag{19}$$

We use the same ideas to treat  $R_2$ . In fact we use the Lipschitz condition on the kernel  $K$  and the assumption (H2) to write that

$$\begin{aligned} |L_1(x_{k(x)}, \bar{y}) - L_1(x_{k(x)}, t_{j(\bar{y})})| &\leq \frac{C}{n\phi(h)\phi(b)} \sum_{i=1}^n K_i(x_{k(x)}) (|K_i(\bar{y}) - k_i(t_{j(\bar{y})})|) \\ &\leq \frac{C}{n} \sum_{i=1}^n Z_i, \end{aligned}$$

where  $Z_i = \frac{w_n K_i(x_{k(x)}) \mathbb{I}_{B(x_k(x),h)}(X_i) \mathbb{I}_{\{Y \leq \bar{y} \leq Y+b\}}}{b\phi(h)\phi(b)}$ .

It is clear that the assumption (H3) permits to write that

$$Z_1 = O\left(\frac{w_n}{b\phi(b)\phi(h)}\right), \mathbb{E}[Z_1] = O\left(\frac{w_n}{b}\right) \text{ and } \text{var}(Z_1) = O\left(\frac{w_n^2}{b^2\phi(b)\phi(h)}\right).$$

Similarly, as previously we get

$$R_2 = O_{a.co.} \left( \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{n\phi(b)\phi(h)}} \right) \text{ and } R_4 = O \left( \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{n\phi(b)\phi(h)}} \right). \tag{20}$$

It remains to evaluate  $R_3$ . Indeed, we write

$$\begin{aligned} P\left(R_3 > \eta \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{n\phi(b)\phi(h)}}\right) &= P\left(\max_{j \in \{1, 2, \dots, d_n\}} \max_{k \in \{1, \dots, N_{\varepsilon}(S_{\mathcal{F}})\}} |L_1(x_k, t_{j(\bar{y})}) - \mathbb{E}L_1(x_k, t_{j(\bar{y})})| > \eta \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{n\phi(b)\phi(h)}}\right) \\ &\leq d_n N_{\varepsilon}(S_{\mathcal{F}}) \max_{j \in \{1, 2, \dots, d_n\}} \max_{k \in \{1, \dots, N_{\varepsilon}(S_{\mathcal{F}})\}} P\left(|L_1(x_k, t_{j(\bar{y})}) - \mathbb{E}L_1(x_k, t_{j(\bar{y})})| > \eta \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{n\phi(b)\phi(h)}}\right) \\ &\leq d_n N_{\varepsilon}(S_{\mathcal{F}}) \max_{j \in \{1, 2, \dots, d_n\}} \max_{k \in \{1, \dots, N_{\varepsilon}(S_{\mathcal{F}})\}} P\left(\left|\frac{1}{n} \sum_{i=1}^n \Gamma_i\right| > \eta \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{n\phi(b)\phi(h)}}\right). \end{aligned}$$

Where

$$\Gamma_i = \frac{1}{\mathbb{E}[K_1(x)K_1(\tilde{y})]} [K_i(x_k)(K_i(t_{j(\tilde{y})}) - E(K_i(x_k)K_i(t_j))].$$

It follows from the fact that the kernel  $K$  is bounded, that  $E|\Gamma_i|^2 \leq C(\phi(b)\phi(h))^{-1}$ . Thus, we apply the Bernstein exponential inequality we obtain for all  $j \leq N_\varepsilon(S_{\mathcal{F}})$ , that

$$P \left( \left| L_1(x_k, t_{j(\tilde{y})}) - \mathbb{E}L_1(x_k, t_{j(\tilde{y})}) \right| > \eta \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{n\phi(b)\phi(h)}} \right) \leq 2 \exp \left\{ -C\eta^2 \Psi_{S_{\mathcal{F}}}(\varepsilon) \right\}.$$

Therefore, by choosing  $C\eta^2 = \beta$ , and using the fact that  $d_n = O(w_n^{-1})$ , we conclude that

$$d_n N_\varepsilon(S_{\mathcal{F}}) \max_{j \in \{1, 2, \dots, d_n\}} \max_{k \in \{1, \dots, N_\varepsilon(S_{\mathcal{F}})\}} P \left( \left| L_1(x_k, t_j) - \mathbb{E}L_1(x_k, t_j) \right| > \eta \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{n\phi(b)\phi(h)}} \right) \leq C' d_n (N_\varepsilon(S_{\mathcal{F}}))^{1-C\eta^2}.$$

Finally, using (H5) and (12) we obtain

$$R_3 = O_{a.co.} \left( \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{n\phi(b)\phi(h)}} \right). \tag{21}$$

By using the same decomposition:

$$\begin{aligned} |L_2(x, y, \tilde{y}) - \mathbb{E}[L_2(x, y, \tilde{y})]| &\leq \underbrace{\sup_{x \in S_{\mathcal{F}}} \sup_{y \in S_{\mathcal{R}}} |L_2(x, y, \tilde{y}) - L_2(x_k, y, \tilde{y})|}_{S_1} \\ &+ \underbrace{\sup_{x \in S_{\mathcal{F}}} \sup_{y \in S_{\mathcal{R}}} |L_2(x_k, y, \tilde{y}) - L_2(x_k, t_{j(y)}, \tilde{y})|}_{S_2} \\ &+ \underbrace{\sup_{x \in S_{\mathcal{F}}} \sup_{y \in S_{\mathcal{R}}} |L_2(x_k, t_{j(y)}, \tilde{y}) - \mathbb{E}[L_2(x_k, t_{j(y)}, \tilde{y})]|}_{S_3} \\ &+ \underbrace{\sup_{x \in S_{\mathcal{F}}} \sup_{y \in S_{\mathcal{R}}} |\mathbb{E}[L_2(x_k, t_{j(y)}, \tilde{y})] - \mathbb{E}[L_2(x_k, y, \tilde{y})]|}_{S_4} \\ &+ \underbrace{\sup_{x \in S_{\mathcal{F}}} \sup_{y \in S_{\mathcal{R}}} |\mathbb{E}[L_2(x_k, y, \tilde{y})] - \mathbb{E}[L_2(x, y, \tilde{y})]|}_{S_5}. \end{aligned}$$

So as before we get:

$$S_1 = O_{a.co.} \left( \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{ng\phi(h)\phi(b)}} \right) \text{ and } S_5 = O \left( \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{ng\phi(h)\phi(b)}} \right). \tag{22}$$

And

$$S_2 = O_{a.co.} \left( \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{ng\phi(b)\phi(h)}} \right) \text{ and } S_4 = O \left( \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{ng\phi(b)\phi(h)}} \right). \tag{23}$$

And

$$S_3 = O_{a.co.} \left( \sqrt{\frac{\Psi_{S_{\mathcal{F}}}(\varepsilon)}{ng\phi(b)\phi(h)}} \right). \tag{24}$$

$$\begin{aligned} &|\mathbb{E}[L_2(x, y, \tilde{y})] - f^x(y)| \\ &\leq C \mathbb{E} \left[ \left| K\left(\frac{d(x, X_1)}{h}\right) K\left(\frac{\tilde{y} - \tilde{Y}_1}{b}\right) \mathbb{E} \left[ \frac{1}{g} K_0 \left( \frac{\tilde{y} - \tilde{Y}_1}{g} \right) - f_Y^{X_1, \tilde{Y}_1}(y) \mid (X_1, \tilde{Y}_1) \right] \right| \right] \\ &\leq C \mathbb{E} \left[ \left| K\left(\frac{d(x, X_1)}{h}\right) K\left(\frac{\tilde{y} - \tilde{Y}_1}{b}\right) \mathbb{E} \left[ \frac{1}{g} K_0 \left( \frac{\tilde{y} - \tilde{Y}_1}{g} \right) \mid (X_1, \tilde{Y}_1) \right] - f_Y^{X_1, \tilde{Y}_1}(y) \right| \right] \end{aligned}$$

Moreover, by change of variable:

$$\mathbb{E} \left[ g^{-1} K_0 \left( \frac{y - Y_1}{g} \right) \mid (X_1, \tilde{Y}_1) \right] = \int_{\mathbb{R}} K_0(u) f_Y^{X_1, \tilde{Y}_1}(y - ug) du.$$

Finally, by (H2) we get:

$$\left| \mathbb{E}[L_2(x, y, \tilde{y})] - f_Y^{X_1, \tilde{Y}_1}(y) \right| = O(g^{\beta_1}). \tag{25}$$

So, The Lemma 5.4 can be easily deduced from (19), (20), (22), (23), (24) and (25). □

### Acknowledgements

The authors greatly thank the Editor in chief and the reviewers for the careful reading, constructive comments and relevant remarks, which permit us to improve the paper.

### References

Carroll, R. J., Knickerbocker, R. K., and Wang, C. Y., (1995). Dimension reduction in a semiparametric regression model with errors in covariates. *The Annals of Statistics*.

Carroll, R. J., Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society*.

Duncan, G. J., Hill, D. H.. (1985). An investigation of the extent and consequences of measurement error in labor-economic survey data. *Journal of Labor Economics*.

Firas, I., Ali Hajj H., and Rachdi, M., (2019). Regression model for surrogate data in high dimensional statistics. *Journal of Communications in Statistics – Theory and Methods*.

Ferraty, F., Laksaci, A., Tadj, A., Vieu, P., (2010). Rate of uniform consistency for non-parametric estimates with functional variables. *Journal of Statistical Planning and Inference*.

- Ferraty, F., Vieu, P., (2006). Nonparametric functional data analysis. Theory and practice, NY: Springer Series in Statistics.
- Ferraty, F., Vieu, P., (2002). The functional nonparametric model and application to spectrometric data. *Comput. Statist.*
- Ferraty, F. Vieu, P., (2011). Kernel regression estimation for functional data. In *The Oxford Handbook of Functional Data Analysis* (Ed. F. Ferraty and Y. Romain). Oxford University Press.
- Hsing, T., Eubank R., (2015). Theoretical foundations of functional data analysis, with an introduction to linear operators. *Wiley series in probability and statistics*. Chichester, UK: John Wiley and Sons.
- Horvath, L., Kokoszka P., (2012). Inference for functional data with applications. New York, NY: Springer Series in Statistics.
- Goia, A., Vieu P., (2016). An introduction to recent advances in high/infinite dimensional statistics. *Journal of Multivariate Analysis*, 146, pp. 1–6.
- Kolmogorov A. N., Tikhomirov V. M., (1959).  $\varepsilon$ -entropy and  $\varepsilon$ -capacity. *Uspekhi Mat. Nauk.*, 14, pp. 3–86., 2, pp. 277–364.
- Lecoutre, J. P., (1990). Uniform consistency of a class of regression function estimators for Banach-space valued random variable. *Statist. Probab. Lett.*
- Loève, M., (1963). *Probability Theory*, 3rd ed. Van Nostranr Princeton.
- Pepe, M. S., (1992). Inference using surrogate outcome data and validation sample. *Biometrika* 79.
- Rachdi, M., Vieu, P., (2007). Nonparametric regression for functional data: Automatic smoothing parameter selection. *Journal of Statistical Planning and Inference*.
- Wang, Q. H., (2000). Estimation of linear error-in-covariables models with validation data under random censorship. *Journal of Multivariate Analysis*.
- Wang, Q. H., Rao, J. N. K., (2002). Empirical likelihood-based in linear errors in covariables models with validation data. *Biometrika*.
- Wang, Q. H., (2003). Dimension reduction in partly linear error-in-response models with validation data. *Journal of Multivariate Analysis*.
- Wang, Q. H., (2006). Nonparametric regression function estimation with surrogate data and validation sampling. *Journal of Multivariate Analysis*.



Wang, J. L., Chiou, J. M., and Muller, H. G., (2016). Review of functional data analysis. *Annual Review of Statistics and Its Application*.

Zhang, J., (2014). Analysis of variance for functional data. *Monographs on Statistics and Applied Probability*.



## Models for survey nonresponse and bias adjustment techniques

H. Öztaş Ayhan<sup>1</sup>

### Abstract

Survey statisticians have been dealing with the issues of nonresponse in sample surveys for many years. Due to the complex nature of the mechanism, so far it has not been easy to find a general solution to this problem. In this paper, several aspects of this topic will be elaborated on: the survey unit nonresponse bias has been examined alternatively by taking response amounts which are fixed initially and also by taking the response amounts as random variables. An overview of the components of the bias due to nonresponse will be performed. Nonresponse bias components are illustrated for each alternative approach and the amount of bias was computed for each case.

**Key words:** response rate, nonresponse bias, nonresponse rate, sample survey, survey error.

### 1. Introduction

During many past studies, the evaluation of the nonresponse bias was based on presenting the nonresponse error in the form of nonresponse rates. However, nonresponse bias may not be related to the response or nonresponse rates of a given study. Increasing response rate (decreasing nonresponse rate) may not always correspond to decreasing nonresponse bias for a given study (Groves and Couper, 1998).

On the other hand, in many studies in the past, the term “bias” was interpreted differently from, how we evaluate the “statistical bias”. The arguments have gone even to far to suggest that, the bias can be obtained within the available inside information on a given sample. However, there are few studies that mentioned statistical bias, which is based on the differences between the “expected value of all possible sample estimates” from the “corresponding parameter” Bethlehem and Kersten (1985), Groves *et al.* (2002), Keser (2011), Kish (1995), Moser and Kalton (1979), Lindström *et al.* (1979), and Lindström (1983). Since, we can only afford to select one sample for a given study,

---

<sup>1</sup> Professor Emeritus, Department of Statistics, METU, Ankara, Turkey. E-mail: oayhan@metu.edu.tr.  
ORCID: <https://orcid.org/0000-0003-3818-483X>.



in this case “the parameter” and “other sample estimates” will also be unknown. Furthermore, by using certain rules during the field operation, the amount of nonresponse can be determined after the fieldwork. Consequently, the amount of nonresponse is only fixed when the field operation is completed.

Alternatively, the amount of nonresponse is unknown before the fieldwork and therefore initially it can be evaluated as a random variable.

The objective of this research is to formulate basic computation of response and nonresponse models. The study also aims to present and discuss alternative response/nonresponse models (fixed response model and random response model). In addition to these, it is aimed to present and compare the alternative bias adjustment techniques for different models.

The impact of nonresponse on the estimators have been examined under two alternative approaches. These are the “fixed response model” and the “random response model” (Lindström *et al.* 1979). Most of the past research is based on assuming that the response and nonresponse amounts are fixed before the survey. Therefore, these studies have used the following nonresponse bias evaluation.

## 2. Taking response amounts which are fixed initially

The fixed response model assumes the population to consist of two mutually exclusive and exhaustive strata: the response stratum and the nonresponse stratum. If selected in the sample, elements in the response stratum will participate in the survey with certainty and elements in the nonresponse stratum will not participate with certainty (Bethlehem, 2009).

The population size of  $N$  can artificially be divided into response and nonresponse stratum. We can use the following form ( $R_i = N_i/N$ ) of the rate and the size

( $N_i = \sum_{j=1}^J N_{ij}$ ) to illustrate the mechanism, where  $i = 1, 2$ .

$$\text{Response rate: } R_1 = N_1/N \text{ and Nonresponse rate: } R_2 = N_2/N \quad (1)$$

$$N_1 + N_2 = N, \quad R_1 + R_2 = 1, \quad (1 - R_1) = R_2 \quad (2)$$

The survey data will only be collected for the response strata. The response strata will have the mean  $\mu_1$  which is based on the  $N_1$  observations.

$$\text{Response stratum mean will be, } \mu_1 = N_1^{-1} \left[ \sum_{j=1}^{N_1} X_{1j} \right]. \quad (3)$$

Nonresponse stratum mean will be, 
$$\mu_2 = N_2^{-1} \left[ \sum_{j=1}^{N_2} X_{2j} \right]. \tag{4}$$

Population mean will be, 
$$\mu = N^{-1} \left[ \sum_{j=1}^N X_j \right]. \tag{5}$$

Household and individual response and nonresponse rates are computed on the basis of the methodology which was proposed by Ayhan (2017).

In a similar way, the sample size of  $n$  can artificially be divided into response and nonresponse stratum. We can use the following form ( $r_i = n_i/n$ ) of the rate and the size ( $n_i = \sum_{j=1}^J n_{ij}$ ) to illustrate the mechanism, where  $i = 1, 2$ .

For **one-stage sample selection**, which can be based on household selection;

Response rate:  $r_1 = n_1/n$  (6)

Nonresponse rate:  $r_2 = n_2/n$  (7)

$n_1 + n_2 = n, \quad r_1 + r_2 = 1, \quad (1 - r_1) = r_2$  (8)

*Household response rate (HRR)* is computed as the ratio of ( $n_1/n$ ) from the selected sample. *Household nonresponse rate (HNRR)* can be taken as the complement of the household response rate (HRR), for first stage sample selection.

$Household\ SurveyRR = \frac{n_1}{n}$  (9)

$HNRR = (n_2/n) = 1 - HRR = [1 - (n_1/n)]$  (10)

For **two-stage sample selection**, which can be based on household survey ( $n_i/n$ ) and individual person ( $m_i/m$ ) selections, as a product;

Household response rate,  $HRR = n_1/n$  (11)

Individual response component,  $IRC = m_1/m$  (12)

Individual response rate,  $IRR = (HRR)(IRC) = (n_1/n)(m_1/m)$  (13)

*Individual nonresponse rate (INRR)* is calculated by the multiplication of *household nonresponse rate* and *individual nonresponse component*. Individual nonresponse component is calculated by taking *nonrespondent individuals ( $m_2$ )* over *enumerated individuals ( $m$ )*.

*Household response rate (HRR)* is computed as the ratio of ( $n_1/n$ ) from the selected sample. *Individual response rate (IRR)* is calculated by the multiplication of household response rates and individual response component. Individual response

component is calculated as respondent individuals ( $m_1$ ) over, enumerated individuals ( $m$ ). These calculations are given with the following formulas.

$$\text{Individual Survey RR} = \frac{n_1 m_1}{n m} \quad (14)$$

$$\text{Individual Survey NRR} = \frac{n_2 m_2}{n m} \quad (15)$$

$$\text{Individual nonresponse component, INRC} = m_2/m \quad (16)$$

$$\text{Individual nonresponse rate, INRR} = (HNRR)(INRC) = (n_2/n)(m_2/m) \quad (17)$$

For two-stage sample selection, individual survey response rate and individual survey nonresponse rate cannot be simple complements.

$$\text{That is, } [(n_1/n)(m_1/m)] \neq 1 - [(n_2/n)(m_2/m)]. \quad (18)$$

When  $n_1$  and  $n_2$  are taken as fixed, where  $n_1 = \frac{N_1}{N} n$  and  $n_2 = \frac{N_2}{N} n$  then

$$E(n_1) E(\bar{x}_1 | n_1) = \frac{N_1}{N} n \quad \text{and} \quad E(n_2) E(\bar{x}_2 | n_2) = \frac{N_2}{N} n \quad (19)$$

Moser and Kalton (1979), Ayhan (1981), and Bethlehem and Kersten (1985) stated that, the bias of nonresponse occurs when the response stratum mean  $\mu_1$  is used instead of the total population mean  $\mu$ . The source of nonresponse bias is based on the use of

$$\text{Lim}_{n \rightarrow N} E(\bar{x}_1) = \mu \quad \text{instead of} \quad \text{Lim}_{n \rightarrow N} E(\bar{x}) = \mu, \quad (20)$$

$$\text{where } \text{Lim}_{n \rightarrow N} E(\bar{x}_1) \neq \mu \quad \text{but} \quad \text{Lim}_{n \rightarrow N} E(\bar{x}_1) = \mu_1. \quad (21)$$

The *nonresponse bias* due to the use of response stratum mean will be,

$$B(\bar{x}_1) = \mu_1 - \mu = \mu_1 - (R_1 \mu_1 + R_2 \mu_2) \quad (22)$$

$$= \mu_1(1 - R_1) - R_2 \mu_2 = R_2(\mu_1 - \mu_2) \quad (23)$$

where  $\mu = (R_1 \mu_1 + R_2 \mu_2)$  and  $(1 - R_1) = R_2$

The effect of bias is based on the *amount of nonresponse rate* and the *difference between the response and nonresponse strata means*. Detailed derivations of the proof are available by Moser and Kalton (1979) and Ayhan (1981).

### 3. Taking response amounts as random variables

Survey nonresponse components and issues of bias have been examined by Bethlehem and Kersten (1985), Bethlehem and Keller (1987) and Bethlehem (2002 & 2009).

The random response model assumes every element in the population to have an unknown response probability. If an element is selected in the sample, a random mechanism is activated that results with a given probability in response and with a complement probability in nonresponse (Bethlehem, 2009).

In order to consider the response amounts as random variables we have proposed the following set of formulations;

$$\text{Define } \mu = \frac{N_1}{N} \mu_1 + \frac{N_2}{N} \mu_2, \quad N = N_1 + N_2 \tag{24}$$

$$\text{Let } \hat{\mu} = \frac{n_1}{n} \bar{x}_1 + \frac{n_2}{n} \bar{x}_2, \quad n = n_1 + n_2 \tag{25}$$

$$E(\hat{\mu}) = \frac{1}{n} E(n_1)E(\bar{x}_1 | n_1) + \frac{1}{n} E(n_2)E(\bar{x}_2 | n_2) \tag{26}$$

$$= \frac{1}{n} \left[ n \frac{N_1}{N} \mu_1 + n \frac{N_2}{N} \mu_2 \right] = \frac{N_1}{N} \mu_1 + \frac{N_2}{N} \mu_2 \tag{27}$$

Putting the two strata together and draw a random sample of size  $n$ . Let  $n_1$  fall into stratum 1, and  $n_2$  fall in stratum 2,  $n_1 + n_2 = n$ .

$$\text{Then } E\left(\frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n}\right) = \frac{n \frac{N_1}{N} \mu_1 + n \frac{N_2}{N} \mu_2}{n} = \frac{N_1 \mu_1 + N_2 \mu_2}{N} = \mu \tag{28}$$

$$E(\bar{x}_1 | n_1) = \mu_1 \quad \text{and} \quad E(\bar{x}_2 | n_2) = \mu_2 \tag{29}$$

When nonresponse occurs at random, it reduces to a single sample situation with sample size  $n$  in which case  $\bar{x}_1$  is estimating  $\mu$ .

There is a real problem with the methodology as follows: Let  $\bar{x}$  be the sample mean.

$$\text{Then, } E(\bar{x}) = \mu \quad \text{or} \quad E(\bar{x}) = \bar{X} \tag{30}$$

$$V(\bar{x}) = \frac{\sigma^2}{n} \quad \text{or} \quad V(\bar{x}) = \frac{S^2}{n} \left(1 - \frac{n}{N}\right) \tag{31}$$

Only if  $n$  is fixed a priori. If  $n$  is a random, the results are different:

$$E(\bar{x}) = E(n) E(\bar{x} | n) \quad (32)$$

$$E(\bar{x}^2) = E(n^2) E(\bar{x}^2 | n) \quad (33)$$

In this case,  $n_1$  (responses) and  $n_2$  (nonresponses) and

$$n_1 + n_2 = n \quad (34)$$

$n$  is the sample size, are both subject to (34). Therefore, the results in (31) are not applicable. In fact,  $n_1$  (or  $n_2$ ) has a binomial distribution with,

$$E(n_1) = n(N_1/N) \quad (35)$$

Here,  $N_1/N$  is the proportion of responses in the population.

The difficulty is that, the value of  $N_1$  is not known. If one estimates  $N_1/N$  by  $n_1/n$ , then  $E(n_1) \cong n(n_1/n) = n_1$  (does not make sense, since the expected value of a random variable cannot be the random variable itself), and that is where the difficulty is. Knowing  $n_1$  a priori which is untenable.

$N_1 + N_2 = N$  Here,  $N_2$  are presumed to be nonresponses;

$$\mu = (N_1/N)\mu_1 + (N_2/N)\mu_2. \quad (36)$$

A sample of size  $n$  is available with  $n_1$  responses (random  $n_1$ ) and  $n_2$  nonresponses;

$$n = n_1 + n_2. \quad (37)$$

Assuming  $N$  is very large and sampling is done without replacement,  $n_1/n$  is a binomial variate with  $E(n_1/n) = N_1/N$  (Tiku, 1964).

$$E\left(\frac{n_1}{n} \bar{x}_1\right) = E\left(\frac{n_1}{n}\right) E(\bar{x}_1/n_1) = \frac{N_1}{N} \mu_1 = \mu - \frac{N_2}{N} \mu_2 \quad (38)$$

$$Bias(\bar{x}_1) = -\frac{N_2}{N} \mu_2 = R_2 \mu_2. \quad (39)$$



If we replace the random variable  $n_1/n$  by its expected value  $N_1/N$  which is mathematically naive,

$$E\left(\frac{n_1}{n} \bar{x}_1\right) \cong E\left(\frac{N_1}{N} \bar{x}_1\right) = \frac{N_1}{N} \mu_1 \tag{40}$$

Consequently,  $E(\bar{x}_1) \cong \mu_1$ ;

$$\begin{aligned} \text{The bias is, } Bias(\bar{x}_1) &= E(\bar{x}_1) - \mu \cong \mu_1 - \left(\frac{N_1}{N} \mu_1 + \frac{N_2}{N} \mu_2\right) \\ &= \frac{N_2}{N} (\mu_1 - \mu_2) = R_2 (\mu_1 - \mu_2); \end{aligned} \tag{41}$$

This is however a very naive approximation.

$$\text{The variance of } \frac{n_1}{n} \bar{x}_1 \text{ is } V\left(\frac{n_1}{n} \bar{x}_1\right) = \frac{N_1}{N} \frac{S_1^2}{n} \left(1 - \frac{N_1}{N}\right). \tag{42}$$

**Remark:** You may notice that, equations (39) and (41) are very different from one another. While equation (39) is mathematically correct, equation (41) is suspicious. The only common ground is when  $N_2/N = 0$ , i.e.,  $N_2 = 0$ , in which case both equations (39) and (41) are equal to zero. Since  $n_1$  is a random variable, the sampling variance of the mean for response stratum is,

$$V(\bar{x}_1) = \frac{S_1^2}{n} \left(1 - \frac{N_1}{N}\right) E\left(\frac{1}{n_1}\right). \tag{43}$$

Thus  $\bar{x}_1$  is not an attractive estimator since  $n_1 = 0$  has to be included in which case the Binomial has to be truncated.

$$\mu = \left(\frac{N_1}{N} \mu_1 + \frac{N_2}{N} \mu_2\right) \text{ and } E\left(\frac{n_1}{n} \bar{x}_1\right) = \frac{N_1}{N} \mu_1 \tag{44}$$

$$\frac{N}{N_1} \frac{n_1}{n} \bar{x}_1 \text{ is an unbiased estimator of } \mu_1. \tag{45}$$

$$E\left(\frac{N}{N_1} \frac{n_1}{n} \bar{x}_1 - \mu\right) \text{ is } \mu_1 - \mu = \mu_1 - \left(\frac{N_1}{N} \mu_1 + \frac{N_2}{N} \mu_2\right) = \frac{N_2}{N} (\mu_1 - \mu_2). \tag{46}$$

$$\text{Bias in } \frac{N}{N_1} \frac{n_1}{n} \bar{x}_1 \text{ is } \frac{N_2}{N} (\mu_1 - \mu_2). \tag{47}$$

#### 4. Bias Adjustment Procedures

A recent research on the survey nonresponse bias adjustment has been proposed by Ayhan (2017). The results have shown the effect of nonresponse and callbacks on the estimation of survey nonresponse bias. The following remedies can also be used to adjust the nonresponse error which has occurred in a given survey. Due to the available means, we cannot elaborate any further for the all possible survey situations.

##### 4.1. Use of auxiliary information from subsampled nonrespondents

The nonresponse bias of the stratum mean estimator is,

$$B(\bar{x}_1) = \mu_1 - (R_1\mu_1 + R_2\mu_2) \quad (48)$$

The design mean can be evaluated as,

$$\hat{\mu} = R_1\bar{x}_1 + R_2\mu_2 \quad (49)$$

Since  $\mu_2$  is not known, the sample estimator will take the following form,

$$\bar{x}_w = \sum_{i=1}^2 R_i x_i = R_1\bar{x}_1 + R_2\bar{x}_2^* \quad (50)$$

where  $\bar{x}_1 = n_1^{-1} \left[ \sum_{j=1}^J X_{1j} \right]$  and  $\bar{x}_2 = n_2^{-1} \left[ \sum_{j=1}^J X_{2j} \right]$  is unknown. By taking a random subsample of size  $m_2$ , a new estimator of the nonresponse stratum mean will take the following form, where  $m_2 = f_b(n_2)$ .

$$\bar{x}_2^* = m_2^{-1} \left[ \sum_{j=1}^J X_{2j} \right] = \hat{\mu}_2 \quad \text{and} \quad E(\bar{x}_2^*) = \mu_2 \quad (51)$$

Here  $f_b$  is the subsampling rate from the nonresponse stratum and can be taken as  $f_b = 0.05$ . The expected value of the subsample estimator will be,  $\lim_{n \rightarrow N} E(\bar{x}_2^*) = \mu_2$ .

On the other hand, the desired estimator of the sample mean is,  $\bar{x} = n^{-1} \left[ \sum_{j=1}^J X_j \right]$

##### 4.2. Domain based weighting adjustments for nonresponse

For the domain-based weighting adjustments for nonresponse, we have proposed the following set of formulations. The probability of selection of the overall sample is

obtained simply by the sampling fraction of the selected sample  $f = x/X = 1/F$  for the total sample. On the other hand, after using some method of stratification, the sampling fraction of any strata is  $f_i = x_i / X_i = 1 / F_i$ .

Design weights (Ayhan 1991 and Verma 1991) for non self-weighting sample designs can be computed for each domain  $i$  with the same probability of selection  $p_i$ .

For a *combined ratio mean*  $\theta = Y/X = \sum_i^H Y_i / \sum_i^H X_i$ , which is estimated by

$$\hat{\theta} = y/x = \sum_i^H y_i / \sum_i^H x_i \quad (52)$$

On the other hand, for a *separate ratio mean*  $\theta_w = \sum_i^H W_i \theta_i$ , estimated by

$$\hat{\theta}_w = \sum_i^H W_i \hat{\theta}_i = \sum_i^H W_i [y_i/x_i] \quad (53)$$

The weight  $W_i = \left[ \frac{\sum_{i=1}^H x_i}{\sum_{i=1}^H \{x_i / [(X/x) p_i]\}} \right] / [(X/x) p_i] = P_0 / P_i$  (54)

where  $\sum_{i=1}^H (W_i x_i) = x$

Here,  $P_0$  has been computed to adjust the overall weighted and unweighted sample to be the same. In addition, a weighting procedure for nonresponse is also essential for self-weighting and nonself-weighting sample design outcomes (Ayhan 2003).

Here  $W_i^* = R_0 / R_i$  where  $R_i = x_i^* / x_i$  is the response rate in domain  $i$ . (55)

The overall response rate ( $R_0$ ) for the design can be computed as,

$$R_0 = \sum_{i=1}^I (W_i x_i) / \sum_{i=1}^I (W_i x_i / R_i) \quad (56)$$

where  $R_0$  is used to adjust the sample sizes to be the same,  $\sum_{i=1}^I (W_i W_i^* x_i) = x$ . (57)

### 5. Conclusions

The evaluation of the nonresponse bias as nonresponse error or nonresponse rate was misleading. The nonresponse bias may seem to be related to the response rates for a given study. Increasing response rate may not always correspond to decreasing

nonresponse bias for a given study. This paper has shown alternative approaches to nonresponse bias. In addition to this, the causes of the nonresponse bias can also be obtained from empirical studies of components and models relating to the covariates of survey participation and non-participation.

The current research examined the response amounts as fixed initially. The proposed methodology has shown the effect of bias of nonresponse which is based as the product of “amount of nonresponse rate” and the “difference between the response and nonresponse strata means” [ $B(\bar{x}_1) = R_2(\mu_1 - \mu_2)$ ].

When the response amounts are taken as random variables, the nonresponse bias has provided the same solution [ $B(\bar{x}_1) = N_2/N(\mu_1 - \mu_2)$ ].

A recent research on the survey nonresponse bias adjustment has been proposed by Ayhan (2017). The current study has examined the nonresponse bias adjustment by using additional auxiliary information from the subsampled nonrespondents. An alternative approach was also used by domain-based weighting adjustments for nonresponse.

## References

- Ayhan, H. Ö., (1981). Sources of Nonresponse Bias in 1978 Turkish Fertility Survey. *Turkish Journal of Population Studies*, 2-3, pp. 104-148.
- Ayhan, H. Ö., (1991). Post Stratification and Weighting in Sample Surveys. *Research Symposium '91*. State Institute of Statistics, Ankara, 11 pp.
- Ayhan, H. Ö., (2003). Combined Weighting Procedures for Post-Survey Adjustment in Complex Sample Surveys. *Bulletin of the International Statistical Institute*, 60(1), pp. 53-54.
- Ayhan, H. Ö., (2017). Effect of Nonresponse and Callbacks on the Estimation of Survey Nonresponse Bias. *Turkish Journal of Population Studies*, 39, pp. 91-107.
- Bethlehem, J. G., Kersten, H. M. P., (1985). On the Treatment of Nonresponse in Sample Surveys. *Journal of Official Statistics*, 1(3), pp. 287-300.
- Bethlehem, J. G., Keller, W. J., (1987). Linear Weighting of Sample Survey Data. *Journal of Official Statistics*, 3(2), pp. 141-154.
- Bethlehem, J. G., (2002). Weighting Nonresponse Adjustments Based on Auxiliary Information. In: Groves, R. M., Dillman, D. A., Eltinge, J. L. & Little, R. J. A. (Eds), *Survey Nonresponse*. New York: John Wiley & Sons.
- Bethlehem, J. G., (2009). *Applied Survey Methods: A Statistical Perspective*. Hoboken, NJ: John Wiley & Sons.

- Groves, R. M., Couper, M. P., (1998). *Nonresponse in Household Interview Surveys*. New York: John Wiley & Sons.
- Groves, R. M., D. A. Dillman, J. L. Eltinge and R. J. A. Little, Eds., (2002). *Survey Nonresponse*. New York: John Wiley & Sons.
- Keser, İ. K., (2011). The History of Survey Sampling. *TurkStat, Journal of Statistical Research*, 8(3), pp. 54-74.
- Kish, L., (1995). *Survey Sampling*. New York: John Wiley & Sons.
- Lindström, H., (1983). *Non-Response Errors in Sample Surveys*. Statistics Sweden, Urval Number 16, 94 pp.
- Lindström, H., J. Wretman, G. Forsman, and Cassel, C., (1979). *Standard Methods for Non-Response Treatment in Statistical Estimation*. National Central Bureau of Statistics, Sweden, 60 pp.
- Moser, C. A., Kalton, G., (1979). *Survey Methods in Social Investigation*. London: Heinemann Educational Books.
- Tiku, M. L., (1964). A Note on the Negative Moments of a Truncated Poisson Variate. *Journal of the American Statistical Association*, 59, pp. 1220-1224.
- Verma, V. K., (1991). *Sampling Methods*. United Nations, Statistical Institute for Asia and the Pacific, Manual for Statistical Trainers, Number 2. Tokyo, Japan.



# Quality adjusted GEKS-type indices for price comparisons based on scanner data

Jacek Bialek<sup>1</sup>

## Abstract

A wide variety of retailers (supermarkets, home electronics, Internet shops, etc.) provide scanner data containing information at the level of the barcode, e.g. the Global Trade Item Number (GTIN). As scanner data provide complete transaction information, we may use the expenditure shares of items as weights for calculating price indices at the lowest (elementary) level of data aggregation. The challenge here is the choice of the index formula which should be able to reduce chain drift bias and substitution bias. Multilateral index methods seem to be the best choice due to the dynamic character of scanner data. These indices work on a whole-time window and are transitive, which is key to the elimination of the chain drift effect. Following what is called an identity test, however, it may be expected that even when only prices return to their original values, the index becomes one. Unfortunately, the commonly used multilateral indices (GEKS, CCDI, GK, TPD, TDH) do not meet the identity test. The paper discusses the proposal of two multilateral indices and their weighted versions. On the one hand, the design of the proposed indices is based on the idea of the GEKS index. On the other hand, similarly to the Geary-Khamis method, it requires quality adjusting. It is shown that the proposed indices meet the identity test and most other tests. In an empirical and simulation study, these indices are compared with the SPQ index, which is relatively new and also meets the identity test. The analytical considerations as well as empirical studies confirm the high usefulness of the proposed indices.

**Key words:** scanner data, product classification, product matching, Consumer Price Index, multilateral indices, GEKS index.

## 1. Introduction

Scanner data have numerous advantages compared to traditional survey data collection because such data sets are much bigger than traditional ones and they contain complete transaction information, i.e. information about prices and quantities at the lowest COICOP (Classification of Individual Consumption by Purpose) level. Scanner data contain expenditure information at the item level (i.e. at the retailer's code or the Global Trade Article Number (GTIN) / European Article number (EAN) / Stock Keeping Unit (SKU) barcode level), which makes it possible to use expenditure shares of items as weights for calculating price indices at the lowest (elementary) level of data aggregation. Most statistical agencies use bilateral index numbers in the CPI measurement, i.e. they use indices which compare prices and quantities of a group of commodities from the current period with the

<sup>1</sup>University of Lodz, Department of Statistical Methods, Lodz, Poland, [jacek.bialek@uni.lodz.pl](mailto:jacek.bialek@uni.lodz.pl) & Statistics Poland, Department of Trade and Services, Poland, [J.Bialek@stat.gov.pl](mailto:J.Bialek@stat.gov.pl).

ORCID: <https://orcid.org/0000-0002-0952-5327>.

© Jacek Bialek. Article available under the CC BY-SA 4.0 licence



corresponding prices and quantities from a base (fixed) period. A multilateral index is compiled over a given time window composed of  $T + 1$  successive months (typically  $T = 12$ ). Multilateral price indices take as input all prices and quantities of the previously defined individual products, which are available in a given time window, i.e. in at least two of its periods. These methods are a very good choice in the case of dynamic scanner data, where we observe a large rotation of products and strong seasonality (Chessa et al., 2017). Moreover, multilateral indices are transitive, which means in practice that the calculation of the price dynamics for any two moments in the time window does not depend on the choice of the base period. By definition, transitivity eliminates the chain drift problem which may occur while using scanner data. The chain drift can be formalized in terms of the violation of the multi period identity test. According to this test, one can expect that when all prices and quantities in a current period revert back to their values from the base period, then the index should indicate no price change and it equals one. Thus, multilateral indices are free from the chain drift within a given estimation time window  $[0, T]$ . Although Ivancic et al. (2011) have suggested that the use of multilateral indices in the scanner data case can solve the chain drift problem, most statistical agencies using scanner data still make use of the monthly chained Jevons index (Chessa et al., 2017).

The Jevons (1865) index is an unweighted bilateral formula and it is used at the elementary aggregation level in the traditional data collection. As the scanner data provide information on consumption, it seems more appropriate to use weighted indices. Unfortunately, bilateral weighted formulas do not take into account all information from the time window, while the frequently chained weighted indices (even superlative) may generate chain drift bias (Chessa, 2015) and therefore do not reflect a reasonable price change over longer time intervals. For this reason, many countries have experimented with multilateral indices or even implemented them for the regular production of price indices (Krsinic (2014), Inklaar and Diewert (2016), Chessa et al. (2017); Chessa (2019), Diewert and Fox (2018), de Haan et al. (2021)).

Following the so-called identity test (International Labour Office, 2004; von der Lippe, 2007), however, one may expect that even when only prices return to their original values and quantities do not, the index becomes one. This test is quite restrictive for multilateral indices and causes some controversy among price statisticians. Nevertheless, it is mentioned among the axioms regarding multilateral indices both in the publications of the European Commission and in journals from the area of official statistics (Zhang et al., 2019). Unfortunately, the commonly used multilateral indices (GEKS, CCDI, GK, TPD, TDH) do not meet the identity test. The main aim of the paper is to present and discuss the proposition of two multilateral indices, the idea of which resembles the GEKS index, but which meet the identity test and most of other axioms. The proposed indices are compared with the multilateral SPQ index method, which is relatively new and also meets the identity test.

## 2. The list of considered multilateral price index methods

Multilateral index methods originate in comparisons of price levels across countries or regions. Commonly known methods include the GEKS method (Gini, 1931; Eltetö and Köves, 1964), the Geary-Khamis (GK) method (Geary, 1958; Khamis, 1972), the CCDI



method (Caves et al., 1982) or the Time Product Dummy Methods (de Haan and Krsinich, 2018). These indices work on the defined time window  $[0, T]$ . The idea of the SPQ multilateral price index is based on the relative price and quantity dissimilarity measure  $\Delta_{SPQ}$  (Diewert, 2020). The price dissimilarity measure is used to link together the bilateral Fisher indices according to the special algorithm, which extends the considered time window in each step.

Before we present the proposed multilateral price indices, let us denote sets of homogeneous products belonging to the same product group in months 0 and  $t$  by  $G_0$  and  $G_t$  respectively, and let  $G_{0,t}$  denote a set of matched products in both moments 0 and  $t$ . Although, in general, the item universe may be very dynamic in the scanner data case, we assume that there exists at least one product being available during the whole time interval  $[0, T]$ . Let  $p_i^\tau$  and  $q_i^\tau$  denote the price and quantity of the  $i$ -th product at time  $\tau$  and  $N_{0,t} = \text{card } G_{0,t}$ .

Since the indices proposed in the work are based on the idea of the GEKS index, let us recall its structure (see Section 2.1).

### 2.1. The GEKS method

Let us consider a time interval  $[0, T]$  of observations of prices and quantities that will be used for constructing the GEKS index. The GEKS price index between months 0 and  $t$  is an unweighted geometric mean of  $T + 1$  ratios of bilateral price indices  $P^{\tau,t}$  and  $P^{\tau,0}$ , which are based on the same price index formula. The bilateral price index formula should satisfy the time reversal test, i.e. it should satisfy the condition  $P^{a,b} \cdot P^{b,a} = 1$ . Typically, the GEKS method uses the superlative Fisher (1922) price index, resulting in the following formula:

$$P_{GEKS}^{0,t} = \prod_{\tau=0}^T \left( P_F^{0,\tau} P_F^{\tau,t} \right)^{\frac{1}{T+1}} . \tag{1}$$

Please note that de Haan and van der Grient (2011) suggested that the Törnqvist price index formula (Törnqvist, 1936) could be used instead of the Fisher price index in the Gini methodology. Following Diewert and Fox (2018), the multilateral price comparison method involving the GEKS method based on the Törnqvist price index is called the CCDI method.

### 3. Axiomatic approach in the multilateral method selection

According to the axiomatic approach, desirable index properties (the so-called “tests”) are defined that a multilateral index may, or may not satisfy. The list of tests for multilateral indices can be found in the guide provided by the Australian Bureau of Statistics (2016) (see the chapter entitled: "CRITERIA FOR ASSESSING MULTILATERAL METHODS"). Interesting considerations concerning tests for price indices in the case of dynamic scanner data sets can be found in Zhang et al. (2019), where the authors - on the basis of the COLI (Cost of Living Index) and COGI (Cost of Goods Index) concepts - focus on five main test for a dynamic item universe (*identity test, fixed basket test, upper bound test, lower bound test and responsiveness test*).

Following the guidelines from the Australian Bureau of Statistics (2016) or the paper by Diewert (2020), we consider a wide set of tests for multilateral indices (see **Appendix A**) assuming that the conditions for their use are met (e.g. a set of matched products over a period of time is never empty).

Please note that the discussed multilateral index formulas (GK, GEKS, CCDI, TPD) meet most of the requirements at the same time, such as the *transitivity*, *multi-period identity test*, *positivity and continuity*, *proportionality*, *homogeneity in prices*, *commensurability*, *symmetry in the treatment of time periods* or *symmetry in treatment of products* tests. However, the discussed indexes differ in terms of the total set of tests they meet. For instance: the GEKS, CCDI and TPD indices do not satisfy the *basket test*, the Geary-Khamis and TPD indices do not satisfy the *responsiveness test to imputed prices* while the GEKS or CCDI can incorporate the imputed prices of missing products, and the *homogeneity in quantities* does not hold in the case of the Geary-Khamis formula. Please also note that the SPQ index is the only multilateral index that satisfies the *identity test*, which is a stronger requirement than the lack of chain drift.

#### 4. Proposition of new multilateral indices

In the "classical" approach to constructing the GEKS-type indices, the bilateral price index formula, which is used in the GEKS' body, is the superlative one. In other words, although the standard GEKS method uses the Fisher indices as inputs (Chessa et al., 2017), other superlative indices are possible choices as well, e.g. the Törnqvist or Walsh indices (van Loon and Roels, 2018; Diewert and Fox, 2018). Moreover, in the paper by Chessa et al. (2017), we can read that "the bilateral indices should satisfy the time reversal test". The choice of the superlative indices as an input for GEKS has its justification in the economic approach, since the superlative indices are considered as to be the best proxies for the Cost of Living Index (International Labour Office, 2004). Please note, however, that the concept of multilateral indices is not based on the COLI framework and requirements for multilateral methods differ from those dedicated to bilateral ones. The *time reversibility* requirement, which allows the GEKS index to be transitive, enables expressing the GEKS index in a more intuitive, quotient form:

$$P_{GEKS}^{0,t} = \prod_{\tau=0}^T \left( \frac{P^{\tau,t}}{P^{\tau,0}} \right)^{\frac{1}{T+1}}. \quad (2)$$

where  $P^{\tau,s}$  is the chosen bilateral price index formula (for  $s = 0, t$ ).

In the next part of the work, two new multilateral indices, the structure of which may resemble the idea of the GEKS index at first glance, were proposed. However, the structure of the base index of the proposed multilateral formulas differs completely from the adopted convention related to the application of the superlative index. Moreover, the calculation of the base index will require quality adjusting, which in turn is more like the Geary-Khamis index idea. In fact, the proposed indices are in a sense a hybrid approach, i.e. they constitute a bridge between the quality adjusted unit value method and the GEKS method.

Finally, it should also be emphasised that one of the proposals (i.e. GEKS-AQU, see

Section 4.1) does not assume that the formula  $P^{\tau,s}$  is a price index, but only a variant of quality adjusted unit value. This is a completely new approach in the theory of multilateral indices but still guarantees good axiomatic properties of the proposed index.

#### 4.1. Proposition based on the asynchronous quality-adjusted unit value

In the unit value concept, prices of homogeneous products are equal to the ratio of expenditure and quantity sold (International Labour Office, 2004; Chessa et al., 2017). However, quantities of different products cannot be added together as in the case of homogeneous products. That is why the idea of quality-adjusted unit value assumes that prices  $p_i^s$  of different products  $i \in G_s$  in month  $s$  are transformed into "quality-adjusted prices"  $\frac{p_i^s}{v_i}$  and quantities  $q_i^s$  are converted into "common units"  $v_i q_i^s$  by using a set of factors  $v = \{v_i : i \in G_s\}$  (Chessa et al., 2017). Thus, the "classical" quality adjusted unit value  $QUV_{G_s}^s$  of a set of products  $G_s$  in month  $s$  can be expressed as follows:

$$QUV_{G_s}^s = \frac{\sum_{i \in G_s} q_i^s p_i^s}{\sum_{i \in G_s} v_i q_i^s} \tag{3}$$

The term "Quality-adjusted unit value method" (QU method for short) was introduced by Chessa (2015; 2016). The QU method is a family of unit value based index methods and its general form can be expressed by the following ratio:

$$P_{QU}^{0,t} = \frac{QUV_{G_t}^t}{QUV_{G_0}^0} \tag{4}$$

In practice, consumer response to price changes can be delayed or even accelerated as consumers not only react to current price changes but also use their own "forecasts" or concerns about future price increases. For example, consumption of thermophilic (seasonal) fruit is likely to be higher in summer because they are cheaper than in winter, when the season is almost over. For instance, some interesting study on "unconventional" consumer behaviour, such as stocking and delayed quantity responses to price changes, and its impact on chain drift bias can be found in the paper by von Auer (2019). Since in practice we often observe prices and quantities that are not perfectly synchronised in time, the following form of the "asynchronous quality-adjusted unit value" is proposed:

$$AQUV_{G_{\tau,s}}^{\tau,s} = \frac{\sum_{i \in G_{\tau,s}} q_i^{\tau} p_i^s}{\sum_{i \in G_{\tau,s}} v_i q_i^{\tau}}, \tag{5}$$

where  $\tau$  is any period from the considered time interval  $[0, T]$ . Obviously it holds that  $AQUV_{G_{s,s}}^{s,s} = QUV_{G_s}^s$ . Let us define now the function  $P^{\tau,s}(v, q^{\tau}, p^{\tau}, p^s)$  as follows:

$$P^{\tau,s}(v, q^{\tau}, p^{\tau}, p^s) = \frac{AQUV_{G_{\tau,s}}^{\tau,s}}{AQUV_{G_{\tau,\tau}}^{\tau,\tau}}. \tag{6}$$

Putting (6) in formula (2) we obtain:

$$P_{GEKS-AQU}^{0,t} = \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,t}} q_i^{\tau} p_i^t}{\sum_{i \in G_{\tau,t}} v_i q_i^{\tau}} \right)^{\frac{1}{T+1}} \cdot \frac{\sum_{i \in G_{\tau,0}} q_i^{\tau} p_i^0}{\sum_{i \in G_{\tau,0}} v_i q_i^{\tau}}. \quad (7)$$

Please note that the proposed index behaves like a GEKS index based on the Laspeyres index in the case of static item universe  $G$ . In fact, if the item universe is static, we obtain

$$\begin{aligned} P_{GEKS-AQU}^{0,t} &= \prod_{\tau=0}^T \left( \frac{\sum_{i \in G} q_i^{\tau} p_i^t}{\sum_{i \in G} v_i q_i^{\tau}} \right)^{\frac{1}{T+1}} = \prod_{\tau=0}^T \left( \frac{\sum_{i \in G} q_i^{\tau} p_i^t}{\sum_{i \in G} q_i^{\tau} p_i^0} \right)^{\frac{1}{T+1}} \\ &= \prod_{\tau=0}^T \left( \frac{\sum_{i \in G} q_i^{\tau} p_i^t}{\sum_{i \in G} q_i^{\tau} p_i^0} \right)^{\frac{1}{T+1}} = P_{GEKS-L}^{0,t}. \end{aligned} \quad (8)$$

Finally, please also note, that theoretically the class of the  $GEKS - AQU$  indices is infinite, since different choices of  $v_i$  factors lead to different index values. We could, for instance, consider  $v_i$  factors defined in the Geary-Khamis multilateral index resulting in a new, hybrid index, which would be a mixture of the GEKS and Geary-Khamis ideas. That would, however, be probably a slow solution. In this paper, we adopt the system of weights  $v_i$  corresponding to the augmented Lehr index (Lamboray, 2017; van Loon and Roels, 2018), where

$$v_i = \frac{\sum_{t=0}^T p_i^t q_i^t}{\sum_{t=0}^T q_i^t}. \quad (9)$$

The following theorem can be proved (see **Appendix B**):

**Theorem 1** *The GEKS-AQU index (7) satisfies the following tests: the transitivity, identity, multi period identity, responsiveness, continuity, positivity and normalisation, price proportionality and weak commensurability. If the item universe is the same in the compared periods 0 and t then the GEKS-AQU index satisfies also the homogeneity in prices and homogeneity in quantities tests.*

#### 4.2. Proposition based on the asynchronous quality-adjusted price index

Let us note that formula (5) can be expressed by using quality-adjusted prices and quantities:

$$AQUV_{G_{\tau,s}}^{\tau,s} = \frac{\sum_{i \in G_{\tau,s}} v_i q_i^{\tau} \frac{p_i^s}{v_i}}{\sum_{i \in G_{\tau,s}} v_i q_i^{\tau}}. \quad (10)$$

If we place all the adjusted prices ( $\frac{p_i^s}{v_i}$ ) with the relative prices ( $\frac{p_i^s}{p_i^\tau}$ ), then we obtain an "asynchronous quality-adjusted price index" (AQI), i.e.

$$AQI_{G_{\tau,s}}^{\tau,s} = \frac{\sum_{i \in G_{\tau,s}} v_i q_i^\tau \frac{p_i^s}{p_i^\tau}}{\sum_{i \in G_{\tau,s}} v_i q_i^\tau} \tag{11}$$

This means that the AQI formula can be treated as a weighted arithmetic mean of partial indices  $\frac{p_i^s}{p_i^\tau}$ , where the weights are proportional to the relative share of the product's adjusted quantities (from the base period  $\tau$ ) in the sum of all adjusted quantities.

In the further part of the work, the GEKS index based on the AQI formula will be marked as GEKS-AQI, i.e. by inserting (11) into the formula (2), we obtain:

$$P_{GEKS-AQI}^{0,t} = \prod_{\tau=0}^T \left( \frac{\frac{\sum_{i \in G_{\tau,t}} v_i q_i^\tau \frac{p_i^t}{p_i^\tau}}{\sum_{i \in G_{\tau,t}} v_i q_i^\tau}}{\frac{\sum_{i \in G_{\tau,0}} v_i q_i^\tau \frac{p_i^0}{p_i^\tau}}{\sum_{i \in G_{\tau,0}} v_i q_i^\tau}} \right)^{\frac{1}{T+1}} \tag{12}$$

Note that the GEKS-AQI index takes into account prices and quantities directly from all time window periods, while the GEKS-AQU index takes into account all quantities but only prices from the reference and base period. However, both formulas indirectly need information about the prices (and quantities) of products from each period in the time window to determine the factors  $v_i$  defined by formula (9). In this way, each new product in the analysed time window has an impact on the final value of the proposed indices.

It is possible to show, analogously to the proofs of Theorem 1 (see **Appendix B**), that the following theorem holds:

**Theorem 2** *The GEKS-AQI index (12) satisfies the following tests: the transitivity, identity, multi period identity, responsiveness, continuity, positivity and normalisation, price proportionality and weak commensurability. If the item universe is the same in the compared periods 0 and t then the GEKS-AQI index satisfies also the homogeneity in prices and homogeneity in quantities test.*

**Remark**

Similarly to the weighted GEKS index (Melser, 2018), it seems to be interesting to consider the following weighted versions of the GEKS-AQU and GEKS-AQI indices:

$$P_{WGEKS-AQU}^{0,t} = \prod_{\tau=0}^T \left( \frac{\frac{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^t}{\sum_{i \in G_{\tau,t}} v_i q_i^\tau}}{\frac{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0}{\sum_{i \in G_{\tau,0}} v_i q_i^\tau}} \right)^{v_\tau} \tag{13}$$

and

$$P_{WGEKS-AQI}^{0,t} = \prod_{\tau=0}^T \left( \frac{\frac{\sum_{i \in G_{\tau,t}} v_i q_i^{\tau} p_i^t}{\sum_{i \in G_{\tau,t}} v_i q_i^{\tau}}}{\frac{\sum_{i \in G_{\tau,0}} v_i q_i^{\tau} p_i^0}{\sum_{i \in G_{\tau,0}} v_i q_i^{\tau}}} \right)^{v_{\tau}}, \quad (14)$$

where the weights concerning the period  $\tau$  can be, for instance, defined as follows:

$$v_{\tau} = \frac{\sum_{i \in G_{\tau}} q_i^{\tau} p_i^{\tau}}{\sum_{\tau=0}^T \sum_{i \in G_{\tau}} q_i^{\tau} p_i^{\tau}}. \quad (15)$$

## 5. Empirical Study

Scanner data from one retail chain in Poland are used in our empirical study, i.e. monthly data on *long grain rice* (subgroup of COICOP 5 group: 011111), *ground coffee* (subgroup of COICOP 5 group: 012111), *drinking yoghurt* (subgroup of COICOP 5 group: 011441) and *white sugar* (subgroup of COICOP 5 group: 011811) sold in 212 outlets during the period from December 2019 to December 2020 (352705 records, which means 210 MB of data). Before price index calculations, the data sets were carefully prepared. First, after deleting the records with missing data and performing the deduplication process, the products were classified into the relevant elementary groups (COICOP 5 level) and, after that, into their subgroups (local COICOP 6 level). The classification process was performed using the `data_selecting()` and `data_classification()` functions from the `PriceIndices` R package (Białek, 2021). The first function requires manual preparation of dictionaries of keywords and phrases that identified individual product groups. The second function was used for problematic, previously unclassified products, and required manual preparation of learning samples based on historical data. The classification itself was based on machine learning techniques using random trees and the XGBoost algorithm (Tianqi and Carlo, 2016). To match products, we used the `data_matching()` function from the `PriceIndices` package. To be more precise: products with two identical codes or one of the codes identical and an identical description were automatically matched. Products were also matched if they had identical one of the codes and the Jaro-Winkler (1989) distance of their descriptions was smaller than the fixed precision value: 0.02. In the last step, just before calculating price indices, two data filters were applied to remove unrepresentative products from the database, i.e. the `data_filtering()` function from the cited package was used. The *extreme price filter* (Białek and Beręsewicz, 2021) was applied to eliminate products with more than a three-fold price increase or more than a double price drop from month to month. The *low sale filter* (van Loon and Roels, 2018) was used to eliminate from the sample products with relatively low sales (almost 30% of products were removed).

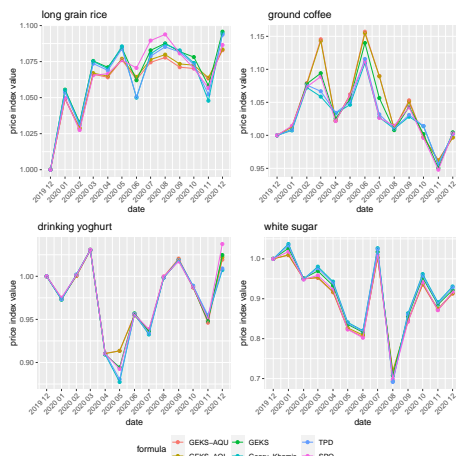


Figure 1: Comparison of selected multilateral indices for four homogeneous groups of food products

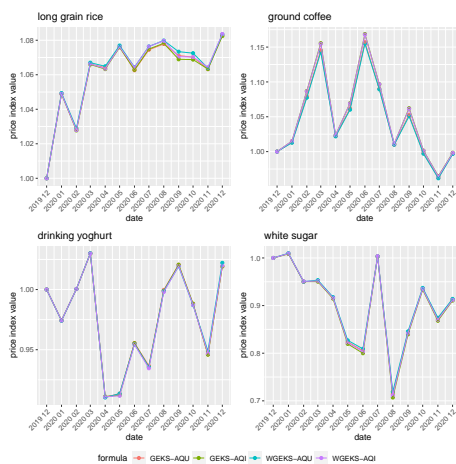


Figure 2: Comparison of the GEKS-AQU and GEKS-AQI indices to their weighted versions for four homogeneous groups of food products

Table 1: Mean absolute differences between considered price indices calculated for **long grain rice**: Dec, 2019 - Dec, 2020 (p.p.)

index	GEKS-AQU	GEKS-AQI	WGEKS-AQU	WGEKS-AQI	GEKS	GK	TPD	SPQ
GEKS-AQU	0.00	0.12	0.06	0.07	0.75	0.87	0.71	0.53
GEKS-AQI	0.12	0.00	0.16	0.08	0.67	0.78	0.63	0.49
WGEKS-AQU	0.06	0.16	0.00	0.09	0.80	0.91	0.76	0.55
WGEKS-AQI	0.07	0.08	0.09	0.00	0.75	0.86	0.71	0.49
GEKS	0.75	0.67	0.80	0.75	0.00	0.29	0.33	0.62
GK	0.87	0.78	0.91	0.86	0.29	0.00	0.15	0.76
TPD	0.71	0.63	0.76	0.71	0.33	0.15	0.00	0.66
SPQ	0.53	0.49	0.55	0.49	0.62	0.76	0.66	0.00

Table 2: Mean absolute differences between considered price indices calculated for **ground coffee**: Dec, 2019 - Dec, 2020 (p.p.)

index	GEKS-AQU	GEKS-AQI	WGEKS-AQU	WGEKS-AQI	GEKS	GK	TPD	SPQ
GEKS-AQU	0.00	0.11	0.54	0.50	1.27	2.39	2.16	1.75
GEKS-AQI	0.11	0.00	0.65	0.62	1.19	2.30	2.07	1.66
WGEKS-AQU	0.54	0.65	0.00	0.04	1.71	2.83	2.58	2.24
WGEKS-AQI	0.50	0.62	0.04	0.00	1.68	2.80	2.56	2.20
GEKS	1.27	1.19	1.71	1.68	0.00	1.30	1.06	0.83
GK	2.39	2.30	2.83	2.80	1.30	0.00	0.27	0.94
TPD	2.16	2.07	2.58	2.56	1.06	0.27	0.00	0.87
SPQ	1.75	1.66	2.24	2.20	0.83	0.94	0.87	0.00

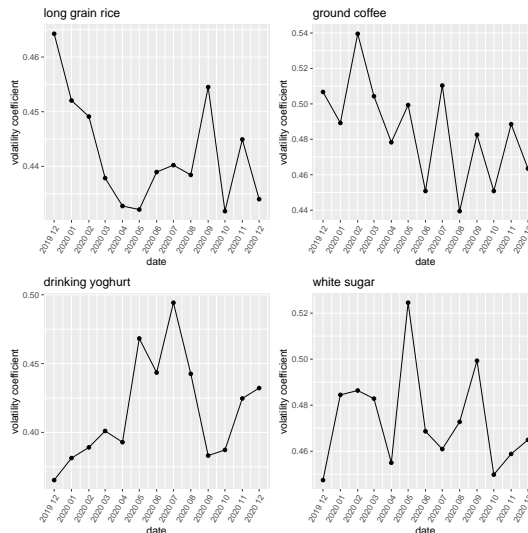


Figure 3: Monthly coefficients of variation of prices calculated for the studied four product groups



Table 3: Mean absolute differences between considered price indices calculated for **drinking yoghurt**: Dec, 2019 - Dec, 2020 (p.p.)

index	GEKS-AQU	GEKS-AQI	WGEKS-AQU	WGEKS-AQI	GEKS	GK	TPD	SPQ
GEKS-AQU	0.00	0.09	0.04	0.08	0.30	0.56	0.50	0.44
GEKS-AQI	0.09	0.00	0.10	0.08	0.25	0.55	0.49	0.40
WGEKS-AQU	0.04	0.10	0.00	0.06	0.30	0.55	0.49	0.45
WGEKS-AQI	0.08	0.08	0.06	0.00	0.28	0.54	0.48	0.44
GEKS	0.30	0.25	0.30	0.28	0.00	0.39	0.36	0.22
GK	0.56	0.55	0.55	0.54	0.39	0.00	0.08	0.52
TPD	0.50	0.49	0.49	0.48	0.36	0.08	0.00	0.47
SPQ	0.44	0.40	0.45	0.44	0.22	0.52	0.47	0.00

Table 4: Mean absolute differences between considered price indices calculated for **white sugar**: Dec, 2019 - Dec, 2020 (p.p.)

index	GEKS-AQU	GEKS-AQI	WGEKS-AQU	WGEKS-AQI	GEKS	GK	TPD	SPQ
GEKS-AQU	0.00	0.21	0.26	0.10	1.24	2.04	1.79	0.50
GEKS-AQI	0.21	0.00	0.47	0.30	1.11	1.92	1.67	0.59
WGEKS-AQU	0.26	0.47	0.00	0.18	1.39	2.19	1.93	0.58
WGEKS-AQI	0.10	0.30	0.18	0.00	1.27	2.08	1.83	0.49
GEKS	1.24	1.11	1.39	1.27	0.00	0.82	0.56	0.99
GK	2.04	1.92	2.19	2.08	0.82	0.00	0.28	1.61
TPD	1.79	1.67	1.93	1.83	0.56	0.28	0.00	1.38
SPQ	0.50	0.59	0.58	0.49	0.99	1.61	1.38	0.00

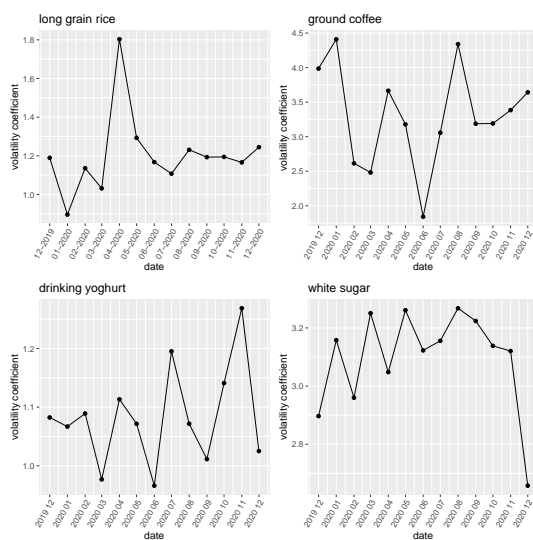


Figure 4: Monthly coefficients of variation of quantities calculated for the studied four product groups

The results obtained for the GEKS-AQU, GEKS-AQI, GEKS, Geary-Khamis, TPD and SPQ indices are presented in Figure 1. An additional comparison of the GEKS-AQU and GEKS-AQI indices to their weighted versions is presented in Figure 2. The average abso-

lute differences between the indices were determined by using the `compare_distances()` function from the `PriceIndices` package (see Tab.1 - Tab.4). Observing Fig.2 and Tab.1-4 we conclude that the largest differences between the GEKS-AQU and GEKS-AQI indices and their weighted versions were observed for data sets: ground coffee (average absolute differences of about 0.5 p.p.) and white sugar (average absolute differences between 0.1 and 0.26 p.p.). In the case of the other two data sets, the corresponding differences turned out to be neglectfully small (they do not exceed 0.09 p.p.). Similarly, the differences between the other multilateral indices appear to be the largest for these ground coffee and white sugar data sets (see Fig.1). As a consequence, there is a suspicion that the distribution of the values of weights based on expenditure shares determines the differences between multilateral indices. However, since the prices of products from homogeneous groups are unlikely to be as diverse as the possible sales levels of these products, there is a natural research hypothesis that the volatility of quantities is the main cause of the differences between multilateral indices. To verify the above-presented hypothesis, an additional analysis was made by determining the monthly coefficients of prices and quantities for all product groups (see Fig.3 and Fig.4).

Price volatility (measured by the coefficient of variation), which is the main cause of differences between bilateral price indices, turned out not to differentiate the analyzed data sets (see Fig.3), and thus it was not price volatility that determined the differences between the values of the indices. As previously suspected, the volatility of the quantity of products sold seems to have a clear impact on the differences between multilateral indices. This conclusion confirms previously obtained results (Białek, 2022). Please note that the coefficients of variation of product quantities are clearly higher for the data sets for white sugar and ground coffee (Fig.4). However, this thread requires further research.

As it was mentioned above, the GEKS-AQU and GEKS-AQI indices approximate each other. Moreover, their values are quite close to those of the GEKS and SPQ indices. The Geary-Khamis index is a good proxy for the Time Product Dummy (TPD) index, which confirms some previous results (Chessa et al., 2017; Białek and Beręsewicz, 2021), but it always seems to be the most distant from the proposed indices.

In terms of how time-consuming their calculations are, the proposed indices seem to be average. More precisely: the GEKS-AQU index requires slightly less computing time than the GEKS-AQI index - in this respect it is better than the TPD or Geary-Khamis indices, but is worse (slower in calculation) than the SPQ or GEKS indices. The last conclusion, however, is not surprising: firstly, the SPQ index does not work on a given time window like other multilateral indices, and secondly, the GEKS index does not perform quality adjustment as the GEKS-AQU and GEKS-AQI indices do.

## 6. Conclusions

The paper proposes two new multilateral indices, the idea of which resembles the GEKS method, but which perform additional quality adjustment and deviate from the classical approach in which the base formula of the GEKS index is a superlative index. The analytical study has confirmed that the two proposed indices (GEKS-AQU and GEKS-AQI) satisfy most of commonly accepted tests for multilateral indices (see Appendix 6) including the

*identity test* (see Theorems 1 and Theorem 2). The empirical study has shown that differences between the proposed indices and other considered multilateral indices appear only with large variability of quantity in homogeneous groups of products. Quite surprisingly, the price volatility did not play a significant role in the empirical study as determinants of differences between multilateral indices (see Section 5). The same study has also shown that the computation time needed in the case of the GEKS-AQU and GEKS-AQI indices is average compared to most other multilateral indices.

It should also be noted that both the previously known multilateral indices (Geary-Khamis, GEKS, TPD, CCDI, and SPQ) as well as the new indices proposed and discussed in the paper (GEK-AQU, GEKS-AQI, and their weighted versions: WGEKS-AQU and WGEKS-AQI) are implemented in the PriceIndices R package (Białek, 2021), and thus the reader can verify their usefulness on their own data sets.

Nevertheless, some aspects of the behaviour of the proposed multilateral indices still remain unexplored. From a practical point of view, it seems interesting, how great the sensitivity of these methods to changing window updating methods is or if the selection of filter thresholds has a considerable influence on the index values. It should be also noted that the issue of choosing between weighted and unweighted versions of a GEKS-type index is not resolved in the literature. On the one hand, the GEKS-type method uses an underlying, bilateral price index that already performs the first weighting. Thus, it seems that additional weighting is an unnecessary waste of time and may even be detrimental due to giving too much weight to leading products and too little weight to products with relatively lower sales. On the other hand, however, the second weighting stage ranks all periods from the time window, whereas the first weighting stage only looks at pairs of periods being compared. Such a dual weighting system can therefore be an alternative to the low sales filter, i.e. it can be considered when one does not choose to pre-filter the data set. From a purely axiomatic point of view, however, it should be borne in mind that unweighted versions of the GEKS index are well recognized in the literature, meanwhile it is uncertain whether the introduction of additional weighting will not affect the set of tests (axioms) satisfied by a given formula.

## Acknowledgements

This publication is financed by the National Science Centre in Poland (grant no. 2017/25/B/HS4/00387). The views expressed are those of the author and not necessarily those of Statistics Poland.

## References

- Australian Bureau of Statistics (2016), *Making Greater Use of Transactions Data to Compile the Consumer Price Index*, Information Paper 6401.0.60.003, Canberra.
- Białek, J. (2021), 'Priceindices – a new R package for bilateral and multilateral price index calculations', *Statistika – Statistics and Economy Journal* **36**(2), 122–141.

- Białek, J. and Beręsewicz, M. (2021), 'Scanner data in inflation measurement: from raw data to price indices', *The Statistical Journal of the IAOS* **37**, 1315–1336.
- Białek, J. (2022), 'Improving quality of the scanner CPI: proposition of new multilateral methods', *Quality and Quantity* **In press**, online at: <https://doi.org/10.1007/s11135-022-01506-6>.
- Caves, D. W., Christensen, L. R. and Diewert, W. E. (1982), 'Multilateral comparisons of output, input, and productivity using superlative index numbers', *Economic Journal* **92**(365), 73–86.
- Chessa, A. (2015), Towards a generic price index method for scanner data in the dutch cpi, in '14th meeting of the Ottawa Group, Tokyo', pp. 20–22.
- Chessa, A. (2016), 'A new methodology for processing scanner data in the Dutch CPI', *Eurostat review of National Accounts and Macroeconomic Indicators* **1**, 49–69.
- Chessa, A. (2019), A comparison of index extension methods for multilateral methods, in 'Paper presented at the 16th Meeting of the Ottawa Group on Price Indices, Rio de Janeiro, Brazil'.
- Chessa, A. G., Verburg, J. and Willenborg, L. (2017), A comparison of price index methods for scanner data, in '15th meeting of the Ottawa Group, Eltville', pp. 10–12.
- de Haan, J., Hendriks, R. and Scholz, M. (2021), 'Price measurement using scanner data: Time-product dummy versus time dummy hedonic indexes', *Review of Income and Wealth* **67**(2), 394–417.
- de Haan, J. and Krsinich, F. (2018), 'Time dummy hedonic and quality-adjusted unit value indexes: Do they really differ?', *Review of Income and Wealth* **64**(4), 757–776.
- de Haan, J. and van der Grient, H. A. (2011), 'Eliminating chain drift in price indexes based on scanner data', *Journal of Econometrics* **161**(1), 36–46.
- Diewert, W. E. (2020), The chain drift problem and multilateral indexes, Technical report, Discussion Paper 20-07, Vancouver School of Economics.
- Diewert, W. E. and Fox, K. J. (2018), 'Substitution bias in multilateral methods for CPI construction using scanner data', *UNSW Business School Research Paper* (2018-13).
- Eltető, O. and Köves, P. (1964), 'On a problem of index number computation relating to international comparison', *Statisztikai Szemle* **42**(10), 507–518.
- Fisher, I. (1922), *The making of index numbers: a study of their varieties, tests, and reliability*, Vol. xxxi, Houghton Mifflin.
- Geary, R. C. (1958), 'A note on the comparison of exchange rates and purchasing power between countries', *Journal of the Royal Statistical Society. Series A (General)* **121**(1), 97–99.

- Gini, C. (1931), 'On the circular test of index numbers', *Metron* **9**(9), 3–24.
- Inklaar, R. and Diewert, W. E. (2016), 'Measuring industry productivity and cross-country convergence', *Journal of Econometrics* **191**(2), 426–433.
- International Labour Office (2004), 'Consumer Price Index Manual: Theory and Practice', Geneva.
- Ivancic, L., Diewert, W. E. and Fox, K. J. (2011), 'Scanner data, time aggregation and the construction of price indexes', *Journal of Econometrics* **161**(1), 24–35.
- Jaro, M. (1989), 'Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida', *Journal of the American Statistical Association* **84**(406), 414–420.
- Jevons, W. S. (1865), 'On the variation of prices and the value of the currency since 1782', *Journal of the Statistical Society of London* **28**(2), 294–320.
- Khamis, S. H. (1972), 'A new system of index numbers for national and international purposes', *Journal of the Royal Statistical Society: Series A (General)* **135**(1), 96–121.
- Krsinich, F. (2014), The FEWS Index: Fixed Effects with a Window Splice–Non-Revisable Quality-Adjusted Price Indexes with No Characteristic Information, in 'Meeting of the group of experts on consumer price indices', pp. 26–28.
- Lamboray, C. (2017), The Geary-Khamis index and the Lehr index: how much do they differ, in 'Paper to be presented at the 15th meeting of the Ottawa Group', pp. 10–12.
- Melser, D. (2018), 'Scanner data price indexes: Addressing some unresolved issues', *Journal of Business and Economic Statistics* **36**(03), 516–522.
- Tianqi, C. and Carlo, G. (2016), Xgboost: A scalable tree boosting system, in 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. ACM', p. 785–794.
- Törnqvist, L. (1936), 'The bank of Finland's consumption price index', *Bank of Finland Monthly Bulletin* pp. 1–8.
- van Loon, K. V. and Roels, D. (2018), Integrating big data in the Belgian CPI, in 'Paper presented at the meeting of the group of experts on consumer price indices, 8-9 May 2018, Geneva, Switzerland'.
- von Auer, L. (2019), The nature of chain drift, in 'Paper presented at the 17th Meeting of the Ottawa Group on Price Indices, 8–10 May 2019, Rio de Janeiro, Brasil'.
- von der Lippe, P. (2007), *Index Theory and Price Statistics*, Peter Lang, Berlin, Germany.
- Zhang, L.-C., Johansen, I. and Nygaard, R. (2019), 'Tests for price indices in a dynamic item universe', *Journal of Official Statistics* **35**(3), 683–697.

## Appendices

### Appendix A: Tests for multilateral indices

Let  $P$  and  $Q$  denote all prices and quantities observed in the time interval  $[0, T]$ , i.e.  $P = [p^0, p^1, \dots, p^T]$ ,  $Q = [q^0, q^1, \dots, q^T]$ , where  $p^t$  and  $q^t$  mean the vector of prices and the vector of quantities of products sold at time  $t$ , respectively. Let us denote by  $P^{0,t}(P, Q)$  the considered multilateral price index defined for the entire time window  $[0, T]$ . The list of commonly accepted tests for that index is as follows:

#### Transitivity

The transitivity means that  $P^{0,t}(P, Q) = P^{0,s}(P, Q)P^{s,t}(P, Q)$  for any  $0 \leq s < t \leq T$ .

#### Identity

This property means that the index equals identity if all prices revert back to their initial level, i.e. if it holds that  $p_i^t = p_i^0$  for  $i \in G_{0,t}$  then  $P^{0,t}(P, Q) = 1$ . We assume here that the item universe is the same at periods 0 and  $t$ .

#### Multi period identity test

This property means that if all prices and quantities revert back to their initial level, the chained index will equal the unity, i.e. if it holds that  $p_i^t = p_i^0$  and  $q_i^t = q_i^0$  for  $i \in G_{0,t}$  then we obtain  $P^{0,1}(P, Q) \times P^{1,2}(P, Q) \times \dots \times P^{t-1,t}(P, Q) = 1$ . We assume here that the item universe is the same at periods 0 and  $t$ .

#### Fixed basket test

If  $G_0 = G_t$  and  $q_i^0 = q_i^t = q_i$  for  $i \in G_{0,t}$ , then  $P^{0,t}(P, Q) = \frac{\sum_{i \in G_{0,t}} p_i^t q_i}{\sum_{i \in G_{0,t}} p_i^0 q_i}$ .

#### Responsiveness test

For  $G_0 \neq G_t$ , if  $p_i^t = p_i^0$  for all  $i \in G_{0,t}$ , then  $P^{0,t}(P, Q)$  cannot always equal one, regardless of sets:  $G_0 \setminus G_t$  and  $G_t \setminus G_0$ .

#### Continuity, positivity and normalization

$P^{0,t}(P, Q)$  is a positive and continuous function of prices and quantities,  $P^{0,0}(P, Q) = 1$ .

#### Price proportionality

If all prices are proportional in the compared periods 0 and  $t$ , i.e.  $p_i^t = k p_i^0$  for all  $i \in G_{0,t}$  and some positive  $k$ , then the price index depends only on this proportion:  $P^{0,t}(P, Q) = k$ . We assume here that the item universe is the same at periods 0 and  $t$ .

#### Homogeneity in quantities

Rescaling the quantities in any  $s$ -th period does not influence the price index, i.e. for any positive  $k$ , it holds that  $P^{0,t}(P, q^0, \dots, k q^s, \dots, q^t) = P^{0,t}(P, q^0, \dots, q^s, \dots, q^t)$ .

**Homogeneity in prices**

Rescaling the prices in the current period changes the price index by the same proportion, i.e. for any positive  $k$ , it holds that  $P^{0,t}(p^0, p^1, \dots, kp^t, Q) = kP^{0,t}(p^0, p^1, \dots, p^t, Q)$ .

**Commensurability**

Changing the units in which prices and quantities are expressed does not change the price index. In other words, if for each time moment  $s \in [0, T]$  we have  $\bar{p}_i^s = \lambda_i p_i^s$  and  $\tilde{q}_i^s = \frac{q_i^s}{\lambda_i}$  for all  $i \in G_s$  ( $\lambda_i > 0$ ), then  $P^{0,t}(\bar{P}, \tilde{Q}) = P^{0,t}(P, Q)$ . If this conditions holds but only for identical values of  $\lambda_i$ , i.e. when  $\lambda_1 = \lambda_2 = \dots \lambda_N = \lambda$ , then the *weak commensurability* is satisfied.

**Appendix B: Proof of Theorem 1**

**Transitivity**

Let us consider such periods  $s$  and  $t$  from the time window  $[0, T]$  that  $0 \leq s < t \leq T$ . We obtain

$$\begin{aligned}
 P_{GEKS-AQU}^{0,s} \times P_{GEKS-AQU}^{s,t} &= \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,s}} q_i^\tau p_i^s}{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0} \right)^{\frac{1}{T+1}} \times \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^t}{\sum_{i \in G_{\tau,s}} q_i^\tau p_i^s} \right)^{\frac{1}{T+1}} = \\
 &= \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^t}{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0} \right)^{\frac{1}{T+1}} = P_{GEKS-AQU}^{0,t}
 \end{aligned}$$

**Identity**

Let us assume that  $G_0 = G_t = G_{0,t}$  and  $p_i^t = p_i^0$  for  $i \in G_{0,t}$ . We have

$$P_{GEKS-AQU}^{0,t} = \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^t}{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0} \right)^{\frac{1}{T+1}} = \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0}{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0} \right)^{\frac{1}{T+1}} = 1.$$

**Multi period identity test**

Let us assume that that  $p_i^t = p_i^0$  and  $q_i^t = q_i^0$  for  $i \in G_{0,t} = G_0 = G_t$ . We obtain

$$\begin{aligned}
 P_{GEKS-AQU}^{0,1} \times P_{GEKS-AQU}^{1,2} \times \dots \times P_{GEKS-AQU}^{t-1,t} &= \\
 &= \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,1}} q_i^\tau p_i^1}{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0} \right)^{\frac{1}{T+1}} \times \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,2}} q_i^\tau p_i^2}{\sum_{i \in G_{\tau,1}} q_i^\tau p_i^1} \right)^{\frac{1}{T+1}} \times \dots \times \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^t}{\sum_{i \in G_{\tau,t-1}} q_i^\tau p_i^{t-1}} \right)^{\frac{1}{T+1}} = \\
 &= \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^t}{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0} \right)^{\frac{1}{T+1}} = \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0}{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0} \right)^{\frac{1}{T+1}} = 1
 \end{aligned}$$

Please note that this proof does not require the condition  $q_i^t = q_i^0$ .

**Responsiveness test**

Let us assume that  $G_0 \neq G_t$  and  $p_i^t = p_i^0$  for all  $i \in G_{0,t}$ . Since  $G_0 \neq G_t$ , we know that for at least one period  $\tau_0$  we have  $G_{\tau_0,t} \neq G_{\tau_0,0} \cap G_{\tau_0,t}$ , for at least one period  $\tau_*$  we have  $G_{\tau_*,0} \neq G_{\tau_*,0} \cap G_{\tau_*,t}$  and, from our initial assumption (see Section 2), we have that  $G_{\tau,0} \cap G_{\tau,t} \neq \emptyset$  for any  $\tau$ . As a consequence, in general we observe  $AQUV_{G_{\tau_0,t}}^{\tau_0,t} \neq AQUV_{G_{\tau_0,t}}^{\tau_0,t}$ , where  $G_{\tau_0,t}^* = G_{\tau_0,0} \cap G_{\tau_0,t}$ . In a similar way we can conclude that  $AQUV_{G_{\tau_*,0}^*}^{\tau_*,0} \neq AQUV_{G_{\tau_*,0}^*}^{\tau_*,0}$ , where  $G_{\tau_*,0}^* = G_{\tau_*,0} \cap G_{\tau_*,t}$ . Thus, from (6), it holds generally that

$$P_{GEKS-AQU}^{0,t} = \prod_{\tau=0}^T \left( \frac{\frac{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^t}{\sum_{i \in G_{\tau,t}} v_i q_i^\tau}}{\frac{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0}{\sum_{i \in G_{\tau,0}} v_i q_i^\tau}} \right)^{\frac{1}{T+1}} \neq \prod_{\tau=0}^T \left( \frac{\frac{\sum_{i \in G_{\tau,0} \cap G_{\tau,t}} q_i^\tau p_i^t}{\sum_{i \in G_{\tau,0} \cap G_{\tau,t}} v_i q_i^\tau}}{\frac{\sum_{i \in G_{\tau,0} \cap G_{\tau,t}} q_i^\tau p_i^0}{\sum_{i \in G_{\tau,0} \cap G_{\tau,t}} v_i q_i^\tau}} \right)^{\frac{1}{T+1}}.$$

Since we assume that prices at compared time moments are identical, i.e.  $p_i^t = p_i^0$  for  $i \in G_{\tau,0} \cap G_{\tau,t}$ , we obtain that

$$P_{GEKS-AQU}^{0,t} \neq \prod_{\tau=0}^T \left( \frac{\frac{\sum_{i \in G_{\tau,0} \cap G_{\tau,t}} q_i^\tau p_i^0}{\sum_{i \in G_{\tau,0} \cap G_{\tau,t}} v_i q_i^\tau}}{\frac{\sum_{i \in G_{\tau,0} \cap G_{\tau,t}} q_i^\tau p_i^0}{\sum_{i \in G_{\tau,0} \cap G_{\tau,t}} v_i q_i^\tau}} \right)^{\frac{1}{T+1}} = 1.$$

**Continuity, positivity and normalisation**

Continuity and positivity are direct consequences of the definition of the GEKS-AQU index. The normalisation property is also an immediate consequence from its form (6), i.e.

$$P_{GEKS-AQU}^{0,0} = \prod_{\tau=0}^T \left( \frac{\frac{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0}{\sum_{i \in G_{\tau,0}} v_i q_i^\tau}}{\frac{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0}{\sum_{i \in G_{\tau,0}} v_i q_i^\tau}} \right)^{\frac{1}{T+1}} = 1.$$

**Price proportionality**

Assumption that the item universe is the same at periods 0 and  $t$  means that  $G_0 = G_t = G_{0,t}$  and also  $G_{\tau,0} = G_{\tau,t}$  for any  $\tau$ . Let us assume that  $p_i^t = k p_i^0$  for all  $i \in G_{0,t}$  and some positive  $k$ . As a consequence, we obtain

$$P_{GEKS-AQU}^{0,t} = \prod_{\tau=0}^T \left( \frac{\frac{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^t}{\sum_{i \in G_{\tau,t}} v_i q_i^\tau}}{\frac{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0}{\sum_{i \in G_{\tau,0}} v_i q_i^\tau}} \right)^{\frac{1}{T+1}} = \prod_{\tau=0}^T \left( \frac{\frac{\sum_{i \in G_{\tau,0}} q_i^\tau k p_i^0}{\sum_{i \in G_{\tau,0}} v_i q_i^\tau}}{\frac{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0}{\sum_{i \in G_{\tau,0}} v_i q_i^\tau}} \right)^{\frac{1}{T+1}} = (k^{T+1})^{\frac{1}{T+1}} = k.$$

**Homogeneity in quantities**

By rescaling the quantities in some  $s$ -th period, we transform a matrix of quantities  $Q$  into the new matrix  $Q^s$ , a vector of quality-adjusting factors  $v$  into the new vector  $v^s$  and we obtain that

$$P_{GEKS-AQU}^{0,t}(P, q^0, \dots, kq^s, \dots, q^t) = \left( \frac{\frac{\sum_{i \in G_{s,t}} k q_i^s p_i^t}{\sum_{i \in G_{s,t}} v_i^s k q_i^s}}{\frac{\sum_{i \in G_{s,0}} k q_i^s p_i^0}{\sum_{i \in G_{s,0}} v_i^s k q_i^s}} \right)^{\frac{1}{T+1}} \times \prod_{\tau=0, \tau \neq s}^T \left( \frac{\frac{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^t}{\sum_{i \in G_{\tau,t}} v_i^\tau q_i^\tau}}{\frac{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0}{\sum_{i \in G_{\tau,0}} v_i^\tau q_i^\tau}} \right)^{\frac{1}{T+1}} =$$

$$= \left( \frac{\frac{\sum_{i \in G_{s,t}} q_i^s p_i^t}{\sum_{i \in G_{s,t}} v_i^s q_i^s}}{\frac{\sum_{i \in G_{s,0}} q_i^s p_i^0}{\sum_{i \in G_{s,0}} v_i^s q_i^s}} \right)^{\frac{1}{T+1}} \times \prod_{\tau=0, \tau \neq s}^T \left( \frac{\frac{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^t}{\sum_{i \in G_{\tau,t}} v_i^\tau q_i^\tau}}{\frac{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0}{\sum_{i \in G_{\tau,0}} v_i^\tau q_i^\tau}} \right)^{\frac{1}{T+1}} = \prod_{\tau=0}^T \left( \frac{\frac{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^t}{\sum_{i \in G_{\tau,t}} v_i^\tau q_i^\tau}}{\frac{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0}{\sum_{i \in G_{\tau,0}} v_i^\tau q_i^\tau}} \right)^{\frac{1}{T+1}}.$$



From the assumption that  $G_0 = G_t$  we conclude that  $G_{\tau,0} = G_{\tau,t}$  and also

$$\begin{aligned} P_{GEKS-AQU}^{0,t}(P, Q^s) &= \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^\tau}{\sum_{i \in G_{\tau,t}} v_i^\tau q_i^\tau} \right)^{\frac{1}{T+1}} = \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^\tau}{\sum_{i \in G_{\tau,t}} v_i^\tau q_i^\tau} \right)^{\frac{1}{T+1}} = \\ &= \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^\tau}{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^\tau} \right)^{\frac{1}{T+1}} = P_{GS-GEKS}^{0,t}(P, Q). \end{aligned}$$

### Homogeneity in prices

By rescaling prices in the current period  $t$ , we transform a matrix of prices  $P$  into the new matrix  $P^t$ , a vector of quality-adjusting factors  $v$  into the new vector  $v^t$  and we obtain that

$$\begin{aligned} P_{GEKS-AQU}^{0,t}(p^0, \dots, kp^t, Q) &= \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,t}} q_i^\tau kp_i^\tau}{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0} \right)^{\frac{1}{T+1}} = \\ &= (k^{T+1})^{\frac{1}{T+1}} \times \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^\tau}{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0} \right)^{\frac{1}{T+1}} = k \times \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^\tau}{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0} \right)^{\frac{1}{T+1}}. \end{aligned}$$

From the assumption that  $G_0 = G_t$  we conclude that  $G_{\tau,0} = G_{\tau,t}$  and also

$$\begin{aligned} P_{GEKS-AQU}^{0,t}(P^t, Q) &= k \times \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^\tau}{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^\tau} \right)^{\frac{1}{T+1}} = k \times \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^\tau}{\sum_{i \in G_{\tau,t}} v_i^\tau q_i^\tau} \right)^{\frac{1}{T+1}} = \\ &= k \times \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^\tau}{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^\tau} \right)^{\frac{1}{T+1}} = k \times P_{GEKS-AQU}^{0,t}(P, Q). \end{aligned}$$

### Commensurability

Let us also notice that the  $P_{GEKS-AQU}^{0,t}$  index fulfils the weak version of the *commensurability* test. In fact, by rescaling  $\tilde{p}_i^s = \lambda p_i^s$  and  $\tilde{q}_i^s = \frac{q_i^s}{\lambda}$  for all  $s \in [0, T]$  and  $i \in G_s$  ( $\lambda > 0$ ), we obtain a new vector of quality adjusting factors  $v^\lambda = \lambda v$  and

$$P_{GEKS-AQU}^{0,t}(\tilde{P}, \tilde{Q}) = \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,t}} \frac{\lambda}{\lambda} q_i^\tau p_i^\tau}{\sum_{i \in G_{\tau,0}} \frac{\lambda}{\lambda} q_i^\tau p_i^0} \right)^{\frac{1}{T+1}} = \prod_{\tau=0}^T \left( \frac{\sum_{i \in G_{\tau,t}} q_i^\tau p_i^\tau}{\sum_{i \in G_{\tau,0}} q_i^\tau p_i^0} \right)^{\frac{1}{T+1}} = P_{GEKS-AQU}^{0,t}(P, Q).$$

Please note, that if  $G_0 = G_t$ , the GEKS-AQU index satisfies the full (strong) version of the *commensurability* test.



## Household expenditure in Africa: evidence of mean reversion

Gbenga A. Olalude<sup>1</sup>, OlaOluwa S. Yaya<sup>2</sup>, Hammed A. Olayinka<sup>3</sup>,  
Toheeb A. Jimoh<sup>4</sup>, Aliu A. Adebisi<sup>5</sup>, Oluwaseun A. Adesina<sup>6</sup>

### Abstract

This paper investigates the mean reversion in household consumption expenditure in 38 African countries; the expenditure series used were the percentage of nominal Gross Domestic Product (GDP), each spanning 1990 to 2018. Due to a small sample size of time series of household expenditure, with possible structural breaks, we used the Fourier unit root test approach, which enabled us to model both smooth and instantaneous breaks in the expenditure series. The results showed non-mean reversion in the consumption expenditure pattern of Egypt, Madagascar and Tunisia, while mean reversion was detected in the remaining 35 countries. Thus, the majority of African countries are on the verge of recession once shocks that affect the growth of GDP are triggered. Findings in this paper are of relevance to policymakers on poverty alleviation programmes in those selected countries.

**Key words:** household expenditure, poverty level, mean reversion, Africa.

JEL Classification: C22; D19; H31

---

<sup>1</sup> Department of Statistics, Federal Polytechnic Ede, Osun State, Nigeria.

E-mail address: [olalude.gbenga@federalpolyede.edu.ng](mailto:olalude.gbenga@federalpolyede.edu.ng). ORCID: <https://orcid.org/0000-0002-9950-0552>.

<sup>2</sup> Economic and Financial Statistics Unit, Department of Statistics, University of Ibadan, Ibadan, Nigeria & Centre for Econometrics and Applied Research, Ibadan, Nigeria. E-mail address: [os.yaya@ui.edu.ng](mailto:os.yaya@ui.edu.ng). ORCID: <https://orcid.org/0000-0002-7554-3507>.

<sup>3</sup> Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, Massachusetts, USA. E-mail address: [haolayinka@wpi.edu](mailto:haolayinka@wpi.edu). ORCID: <https://orcid.org/0000-0002-9796-5276>.

<sup>4</sup> Department of Mathematics and Statistics, University of Limerick, Ireland.

E-mail address: [toheeb.jimoh@ul.ie](mailto:toheeb.jimoh@ul.ie). ORCID: <https://orcid.org/0000-0003-3830-7641>.

<sup>5</sup> Environmental Statistics Unit, Department of Statistics, University of Ibadan, Ibadan, Nigeria. E-mail address: [aadebiyi083@stu.ui.edu.ng](mailto:aadebiyi083@stu.ui.edu.ng). ORCID: <https://orcid.org/0000-0002-6340-098X>.

<sup>6</sup> Department of Statistics, Ladoko Akintola University of Technology, Ogbomosho, Nigeria. E-mail address: [oaadesina26@lautech.edu.ng](mailto:oaadesina26@lautech.edu.ng). ORCID: <https://orcid.org/0000-0001-6952-3991>.

© G. A. Olalude, O. S. Yaya, H. A. Olayinka, T. A. Jimoh, A. A. Adebisi, O. A. Adesina. Article available under the CC BY-

## 1. Introduction

One of the Sustainable Development Goals (SDGs) is to eliminate extreme poverty in any region of the globe by 2030, and as at the time of writing this paper, poverty is still prevalent in most parts of Africa. Many rich African countries, for instance, South Africa, Nigeria, and Senegal still have quite a high proportion of poor people. Household consumption expenditure amounts to about 60% of the Gross Domestic Product (GDP) as it includes government transfers, the total amount spent by household residents in catering for needs such as clothing, food, housing, transportation, etc., as well as other miscellaneous services that directly benefit households (OECD, 2020).

In 2012, the African Development Bank reported that the four largest economies in Africa in order as South Africa, Egypt, Nigeria, and Algeria accounted for 55% of the overall African household final consumption expenditure for 2009, and of the four, Nigeria is referred to as only a middle-expenditure country. Also, household expenditure varies generally among African countries concerning basic needs such as transportation, food, etc., and their areas, be it urban or rural. For instance, urban households allocate a greater percentage of their expenditures to housing compared to rural areas, except in Uganda, Malawi, and Ethiopia (Lozano-Gracia and Young, 2014).

Based on the latest household consumption expenditure dataset (2018), Africa has the highest average (84.92) household consumption expenditure.<sup>7</sup> Africa also has the second-largest population in the world, after the Asia continent. All these reasons enliven our interest in studying the time-series dynamics of household consumption expenditure across African countries.

Mean reversion is used as a financial term for the assumption that the asset price and historical returns tend to revert to the mean price over time (Mahdavi, 2013). Based on this extract, the strategies of mean reversion on household expenditures work on the assumption that there is an underlying fixed trend in the expenditures of households, and the expenditure of a particular household is assumed to revert to its previous state from the long-term norm. Mean reversion of household expenditure can take place in two folds: firstly, the expected expenditure of household can go in a direction opposite to that of the expenditure; and secondly, the expected expenditure can revert toward a mean level. Reverting to mean level in the sense of a stationary series with supposed constant mean as in Box et al. (2015). Non-mean reversion in the case of nonstationary series, which deviates totally from the mean level of the series. As posited by Ben (2015), as households progress through life, they experience differing successes. Some

---

<sup>7</sup> Europe: 58.35; North America: 67.88; Asia: 56.10; Africa: 84.92, South America: 68.23; and Australia: 54.65 (these were computed by the authors based on datasets available on [databank.worldbank.org/world-development-indicators](http://databank.worldbank.org/world-development-indicators)).

households generate high incomes, however, others generate less. For many households, there is evidence of income fluctuation throughout life, which shows both upward and downward trends.

The fluctuation of household income, therefore, influences the welfare and the behavioural pattern of the households. Ben ascertained that mean reversion is a crucial measure of high-frequency household expenditure. The ability to understand the household spending patterns particularly among the poor would be useful to manage their expenditure as preparation for any change in the economic uncertainties that are expected in the future (Nik et al., 2014). The fact that, at the time of writing this paper, there is no much review on unit roots or mean reversion on household income and expenditure, raises the dire need of embarking on the research for the growth of the African economy. However, a few pieces of literature are cited to have a broad overview of the nature of this research work.

Thankgod (2014) used Keynes' hypothesis of absolute income and found evidence of a long-run association between the variables, hence gave room for estimating a parsimonious error correction model. This in turn indicated a positive relationship between the national income and private consumption expenditure. Akhand (2011) examined the consumption behaviour of households in Indonesia with the use of Friedman's permanent income hypothesis under rational expectation. The unit root test results suggested that the real household consumption per-capita and the real disposable income per-capita in Indonesia follow a random walk process, and hence are eligible to form a cointegrating relationship.

Giray (2013) used Heterogeneous panel unit root tests and the Modified ADF unit root test to investigate the stochastic characteristics of the income-consumption ratios of eleven countries in Central and Eastern Europe (CEE). The study findings foresaw significant evidence of the existence of a hypothesis that indicated that the income-consumption ratio is mean-reverting. Ben (2015) used household consumption and wealth, as well as UK income panel data, and concluded that it is crucial and necessary to account for the mean reversion of shocks in the construction of life-cycle consumption models.

The present paper investigates the mean reversion in household consumption expenditure in 38 African countries, spanning across the four regions of the continent. The analysis is based on the ADF test with Fourier nonlinearity with smooth and instantaneous breaks (FADF-SB) as proposed in Furuoka (2017). The approach works quite well, particularly due to the small sample size of expenditure series since unit root lag is fixed to unity. Three other tests: the ADF, FADF, and ADF-SB are restricted tests to the FADF-SB test. In selecting the best representative test regression model, Furuoka (2017) set out an F-test strategy, which is adopted and applied in this paper. This univariate Fourier-based ADF test, and its panel version based on Seemingly Unrelated

Regressions have been widely applied in Fumitaka et al. (2020), Awolaja et al. (2021), Gil-Alana and Yaya (2021), Yaya (2018), Yaya et al. (2019a,b), Yaya et al. (2020). Yaya et al. (2021), among others.

The remaining part of this paper is structured as follows: Section 2 presents the data and unit root tests employed with the F-tests. Section 3 of the paper presents the empirical findings while Section 4 renders the concluding remark.

## 2. Data and Methods

The data used in this paper are the annual time series of average final consumption expenditure of households in Africa computed as a percentage of nominal GDP. These were obtained from the World Development Indicators (WDI) of the World Bank at the website: <https://data.worldbank.org/indicator/>. Household expenditure series of thirty-eight (38) African countries were included, and each time series spanned between 1990 and 2018.

Table 1 presents the summary report of the dataset, showing the household expenditures across the countries in 1990 and 2018. The table includes the minimum and maximum expenditures in the periods sampled throughout the considered countries. The table also includes the rank of countries in ascending order concerning their household consumption expenditure changes between 1990 and 2018. The ranking was done based on the difference between the household expenditure of 1990 and 2018 for each country. The country with the highest positive difference was ranked as the 1st (Nigeria, 46.39), while the country with the highest negative difference was ranked as the 38th (Chad, -26.77). Botswana, Gabon, Congo, Algeria, Cameroon, Guinea, Mauritius, Morocco, Namibia, Seychelles, South Africa, and Tunisia have expenditure rate ranges from about 51% to 80%. Burundi, Chad, Tanzania, and Uganda have a very high expenditure rate above 100% of their GDP. As of 2018, the expenditure rates of Congo, Gabon, Tanzania, and Uganda dropped within the ranges from about 2% to 18%. Central African Republic, Sierra Leone, and the Gambia improved with a rate that ranges from about 6% to 22%. The range between the minimum and maximum rates varied widely across all the countries considered. This implies that there are fluctuations in household expenditure rates across the years sampled, and these imply high expenditure rates in Africa.

**Table 1:** Data Summary

Country	Code	1990 Exp.	2018 Exp.	Min. Exp.	Max. Exp.	Rank
Algeria	DZA	72.90	59.61	42.94	73.44	32nd
Benin	BEN	93.30	83.03	83.03	96.64	25th
Botswana	BWA	57.37	68.26	55.77	72.60	9th
Burkina Faso	BFA	94.58	81.72	79.19	100.09	31st
Burundi	BDI	105.37	104.12	97.84	113.78	20th
Cameroon	CMR	79.32	81.59	76.92	83.46	16th
Central Afr. Republic	CAF	96.10	102.19	89.91	102.19	14th
Chad	TCD	107.67	80.90	69.19	140.81	38th
Congo	COG	51.36	38.98	35.07	91.45	30th
Congo, Dem. Rep.	COD	88.21	77.89	72.46	101.00	26th
Egypt	EGY	83.30	93.80	82.89	98.22	10th
Gabon	GAB	63.14	49.19	39.51	65.12	34th
Ghana	GHA	94.53	81.01	75.81	102.96	33rd
Guinea	GIN	77.84	96.06	77.61	105.29	3rd
Guinea-Bissau	GNB	97.17	96.23	92.99	110.06	19th
Kenya	KEN	81.03	94.68	77.57	95.69	6th
Madagascar	MDG	94.45	85.52	80.20	100.34	24th
Malawi	MWI	86.60	96.43	85.80	104.11	11th
Mali	MLI	95.58	90.17	82.93	100.04	22nd
Mauritania	MRT	95.12	83.43	58.16	100.23	28th
Mauritius	MUS	76.99	90.91	73.07	90.91	5th
Morocco	MAR	74.45	77.00	74.28	80.97	15th
Mozambique	MOZ	93.60	87.44	82.72	109.44	23rd
Namibia	NAM	76.86	93.62	76.86	97.25	4th
Niger	NER	95.75	80.43	80.29	99.95	35th
Nigeria	NGA	35.79	82.18	35.79	86.92	1st
Rwanda	RWA	93.80	92.40	90.70	148.51	21st
Senegal	SEN	97.64	85.52	85.52	99.09	29th
Seychelles	SYC	79.69	86.28	70.32	86.28	13th
Sierra Leone	SLE	86.30	108.21	86.30	120.16	2nd
South Africa	ZAF	79.91	81.21	78.67	82.62	17th
Sudan	SDN	90.30	78.73	74.40	95.72	27th
Tanzania	TZA	100.56	83.04	67.94	103.15	37th
The Gambia	GMB	89.34	102.69	87.14	105.45	7th
Togo	TGO	85.29	85.63	68.44	100.23	18th
Tunisia	TUN	79.98	92.72	76.49	93.05	8th
Uganda	UGA	100.48	84.30	81.94	100.77	36th
Zimbabwe	ZWE	82.55	90.03	78.19	121.46	12th

**Note:** Rates are given in percentages of Nominal GDP.

The ADF unit root test involving three regression specifications namely: (i) no intercept and trend (ii) intercept only and (iii) intercept and trend, was carried out and the results of the test were contained in Table 2.

Furthermore, the automatic selection of augmented lags was examined, thereafter used the minimum Schwarz information criteria in selecting the optimal lag, and is contained in squared brackets. Noting that these optimal lags may be large enough to bias the unit root decisions in some cases. In the case of no intercept, the null hypothesis of a unit root in household consumption expenditure series was not rejected in virtually all the countries except in Uganda. In this sense, rejection of unit root implies mean reversion evidence, while acceptance of null of unit root implies non-mean reversion in the time series. Due to the magnitude of the time series, the test regression model with no intercept would have under-represented the unit root decision. Meanwhile, by considering models with intercept only, and intercept with time trend, the authors found improved results.

The two-unit root regression models jointly determined unit roots in household consumption expenditure series of Cameroun, Guinea-Bissau, Malawi, Rwanda, Seychelles, Sierra Leone, Gambia, and Togo (21%). Altogether, unit root regression with intercept and trend detected unit root in household expenditure series of Benin, Cameroon, Chad, Egypt, Guinea-Bissau, Malawi, Niger, Nigeria, Rwanda, Senegal, Seychelles, Sierra Leone, Sudan, Gambia, Togo, and Uganda (42.1%). The test regression model with only the constant detected unit root in Burundi, Cameroon, Congo, Congo DR., Gabon, Guinea-Bissau, Malawi, Rwanda, Seychelles, Sierra Leone, Gambia, and Togo (approximately, 31.6%).

These inconsistencies in the ADF decision occurred due to the small size of the series – in this particular case a sample size of 29. Also, it is necessary to care for inherent structural breaks during the unit root test. We described below the unit root frameworks that cater to these shortcomings.



**Table 2:** Results of ADF Unit root tests

Country	Code	None	Intercept	Intercept and trend
Algeria	DZA	-0.7815 [0]	-1.7534 [0]	-1.4175 [0]
Benin	BEN	-0.9044 [0]	-1.7888 [0]	<b>-3.6599 [1]</b>
Botswana	BWA	0.3293 [0]	-2.4362 [0]	-2.6255 [0]
Burkina Faso	BFA	-0.7990 [0]	-1.5384 [0]	-2.0962 [0]
Burundi	BDI	-0.1629 [0]	<b>-3.6057 [0]</b>	-3.5479 [0]
Cameroon	CMR	0.4097 [1]	<b>-3.3211 [0]</b>	<b>-4.1486 [0]</b>
Central Afr. Republic	CAF	0.2679 [0]	-2.7199 [0]	-3.3771 [0]
Chad	TCD	-0.7082 [1]	-2.9566 [0]	<b>-3.9549 [0]</b>
Congo	COG	-0.1139 [2]	<b>-3.4651 [1]</b>	-3.4094 [1]
Congo, Dem. Rep.	COD	-0.6807 [3]	<b>-4.6115 [0]</b>	-3.2890 [2]
Egypt	EGY	0.9748 [0]	-0.9641 [0]	<b>-3.6447 [5]</b>
Gabon	GAB	-0.9538 [2]	<b>-2.9961 [0]</b>	-3.1788 [0]
Ghana	GHA	-0.6397 [0]	-1.5135 [0]	-1.8238 [0]
Guinea	GIN	0.5797 [1]	-2.0094 [0]	-3.3989 [0]
Guinea-Bissau	GNB	-0.1155 [1]	<b>-4.0749 [0]</b>	<b>-3.9617 [0]</b>
Kenya	KEN	0.9416 [0]	-1.7497 [0]	-1.6344 [0]
Madagascar	MDG	-0.5678 [0]	-1.9170 [0]	-3.2636 [0]
Malawi	MWI	0.2770 [1]	<b>-4.0283 [0]</b>	<b>-4.0293 [0]</b>
Mali	MLI	-0.4038 [0]	-1.9220 [0]	-1.8934 [0]
Mauritania	MRT	-0.5427 [0]	-2.7786 [0]	-3.0788 [0]
Mauritius	MUS	1.3613 [0]	-0.0905 [0]	-2.3536 [0]
Morocco	MAR	0.2441 [0]	-2.5937 [0]	-2.6143 [0]
Mozambique	MOZ	-0.3823 [0]	-1.5728 [0]	-2.1244 [0]
Namibia	NAM	0.5374 [0]	-2.8040 [0]	-3.2824 [0]
Niger	NER	-1.6206 [1]	-0.8629 [1]	<b>-4.0671 [0]</b>
Nigeria	NGA	3.4296 [5]	-0.1365 [5]	<b>-4.8912 [0]</b>
Rwanda	RWA	-0.3431 [0]	<b>-3.6677 [0]</b>	<b>-4.8875 [0]</b>
Senegal	SEN	-1.8074 [0]	-1.1199 [0]	<b>-4.1930 [6]</b>
Seychelles	SYC	-0.0200 [2]	<b>-5.2911 [0]</b>	<b>-5.1670 [0]</b>
Sierra Leone	SLE	0.6118 [1]	<b>-3.3522 [0]</b>	<b>-3.6541 [0]</b>
South Africa	ZAF	0.2581 [0]	-2.3954 [1]	-2.7995 [1]
Sudan	SDN	-0.8751 [2]	-2.2498 [0]	<b>-3.6334 [0]</b>
Tanzania	TZA	-1.0607 [0]	-2.1177 [2]	0.2001 [2]
The Gambia	GMB	0.2901 [1]	<b>-4.8758 [0]</b>	<b>-4.7726 [0]</b>
Togo	TGO	-0.2193 [0]	<b>-4.3405 [0]</b>	<b>-4.6638 [0]</b>
Tunisia	TUN	1.7287 [0]	1.0535 [0]	-1.1212 [0]
Uganda	UGA	<b>-2.0633 [2]</b>	-2.3776 [2]	<b>-3.6275 [0]</b>
Zimbabwe	ZWE	0.2162 [1]	-1.8254 [0]	-0.7937 [1]

**Note:** Bolded figures denote that the ADF test is significant at 5% level, and reported in square brackets is the optimal lag length of the augmentation.

The traditional ADF unit root test does not account for structural breaks, in the long run, household expenditure rate can however experience smooth or instantaneous breaks within the considered years (see Perron, 1989; Furuoka, 2017a). As in Enders and Lee (2012a,b), to account for the limitation of the traditional ADF test, they expanded the traditional ADF test to a nonlinear framework with the use of a Fourier function with different frequencies. The general equation of the Fourier form is given as:

$$G(t) = \alpha + \beta t + \sum_{j=1}^m \lambda_j \sin\left(\frac{2\pi jt}{N}\right) + \sum_{j=1}^m \gamma_j \cos\left(\frac{2\pi jt}{N}\right); m \leq \frac{N}{2}; t = 1, 2, \dots \quad (1)$$

where  $\alpha$  and  $\beta$  represent the model intercept and coefficient of the trend, respectively;  $\lambda_j$  and  $\gamma_j$  are the measures of the amplitude and displacement of the sinusoidal component of the deterministic term, respectively;  $\pi$  is canonically taken to be 3.1416;  $m$  is the optimal number of frequencies, and it is to be obtained by the information criteria;  $j$  is a specific frequency, which is set to 1, 2, ..., up to  $m$  initially; and  $N$  represents the total number of observations – the length of the household expenditure rate in this paper.  $\lambda_j$  and  $\gamma_j$  are the nonlinear parameters in the Fourier function that was set up and are assumed to be real values upon estimation. The entire function in (1) becomes a linear function if the values of  $\lambda_j$  and  $\gamma_j$  are 0, therefore, the significance of at least one of  $(\lambda_j, \gamma_j)$  indicates nonlinearity. The classical ADF test regression is given as:

$$\Delta Exp_t = \alpha + \beta t + (\rho - 1)Exp_{t-1} + \sum_{i=1}^p d_i \Delta Exp_{t-i} + \varepsilon_t \quad (2)$$

where  $Exp_t$  is the household expenditure rate specific to a country at the time  $t$ ;  $\varepsilon_t$  is the error term;  $\rho$  is the slope parameter specific to the first lagged dependent variable;  $Exp_{t-1}$  is 1, when the series contains unit root attributes;  $d$  and  $p$  represent the slope and the lag length for the augmentation in the augmented component, respectively. Putting equation (2) and (1) together resulted in the Fourier ADF (FADF) test regression as developed by Enders and Lee's:

$$\Delta Exp_t = \alpha + \beta t + (\rho - 1)Exp_{t-1} + \sum_{j=1}^m \lambda_j \sin\left(\frac{2\pi jt}{N}\right) + \sum_{k=1}^n \gamma_k \cos\left(\frac{2\pi kt}{N}\right) + \sum_{i=1}^p d_i \Delta Exp_{t-i} + \varepsilon_t \quad (3)$$

While testing for a unit root in a given time series modelling, the FADF unit root test accounts for smooth breaks (Becker et al., 2006). Furuoka (2017a), extended the test with a structural break obtained simultaneously in the process. This process aligns with Perron's (2006) one structural break unit root test. Hence, in this study, both the ADF-SB as in Perron (2006) and the FADF-SB as in Furuoka (2017a), are utilised respectively and given as:

$$\Delta Exp_t = \alpha + \beta t + \delta DU_t + \theta D(N_B)_t + (\rho - 1)Exp_{t-1} + \sum_{i=1}^p c_i \Delta Exp_{t-i} + \varepsilon_t \tag{4}$$

$$\Delta Exp_t = \alpha + \beta t + \delta DU_t + \theta D(N_B)_t + (\rho - 1)Exp_{t-1} + \sum_{k=1}^n \lambda_k \sin\left(\frac{2\pi kt}{N}\right) + \sum_{k=1}^n \gamma_k \cos\left(\frac{2\pi kt}{N}\right) + \sum_{i=1}^p c_i \Delta Exp_{t-i} + \varepsilon_t \tag{5}$$

where  $\delta$  represents the coefficient of the structural break dummy variable  $DU_t$ , where  $DU_t = 1$  if  $t > N_B$ , otherwise,  $DU_t = 0$ ;  $N_B$  denotes the break date;  $\theta$  represents the coefficient of the one-time break dummy, where  $D(N_B) = 1$  if  $t = N_B$ ,  $D(N_B) = 0$  otherwise. As the same with the ADF test, the null hypothesis of unit root,  $\rho - 1 = 0$  was tested using a t-test, in the above models represented as the equation (3), (4), and (5). These correspond to FADF, ADF-SB, and FADF-SB unit root tests respectively. The optimal frequency  $\hat{j}$  in equations (3) and (5) is obtained by reducing the residual sum of squares errors (SSR) to its minimum value through

$$SSR_{FADF}(\hat{j}) = \inf_j SSR_{FADF}(j); \quad SSR_{FADF-SB}(\hat{j}) = \inf_j SSR_{FADF-SB}(j) \tag{6}$$

whereas considering ADF-SB and FADF-SB cases, as shown in Perron (2006) and Zivot and Andrews (1992), one structural break is determined endogenously, and not exogenously. The  $(\hat{N}_B)$ , structural break date, is then obtained. Furuoka (2014; 2017a) suggested the use of an F-statistic,

$$F = \frac{(SSR_0 - SSR_1)/k}{SSR_1/(N-r)}, \tag{9}$$

where  $SSR_1$  denotes the unrestricted model sum of squares residuals (SSR);  $SSR_0$  denotes the restricted model of SSR;  $k$  represents the number of restrictions present in the restricted model, and  $r$  represents the number of regressors contained in the unrestricted model. For clarity's sake, the FADF model is an unrestricted model of the ADF regression model in a case when the nonlinear trigonometrical terms are zeros, that is,  $\lambda_j = \gamma_j = 0$ . Also, the ADF-SB model is an unrestricted model of the ADF model when there is no structural break observed. Moreover, the FADF-SB model is an unrestricted model of the ADF model in a case where structural break and nonlinearity forms are not included in the model. Furthermore, the FADF-SB regression is an unrestricted model to the FADF model when the structural break dummies in the model are not found. Finally, the FADF-SB model is an unrestricted model to the ADF-SB regression in a case whereby nonlinearity form via trigonometry is not included. Hence, there are cases of five pairings considered, given as  $F_{FADF\_ADF}$ ,  $F_{ADF-SB\_ADF}$ ,  $F_{FADF-SB\_ADF}$ ,  $F_{FADF-SB\_FADF}$  and  $F_{FADF-SB\_ADF-SB}$  tests. The critical values for each pairing can be found in Furuoka (2017a). Considering the pairing cases, in a case where

there is no significant improvement in an unrestricted model against a restricted one, the model which contains the lowest value of Type I error was accepted to be a better model. The accepted model determined the acceptance of the hypothesis of the unit root of the household expenditure rate.

### 3. Empirical Findings

Concerning the unit root approach described above, the ADF test, whereby the augmentation lag fixed to unity (i.e.  $p = 1$ ) was conducted, and the same augmentation lag was fixed for the other tests, ADF-SB, FADF, and FADF-SB. The results of the unit root tests are presented in Table 3; and the result of the robustness test using the F test is in Table 4. Based on the ADF test result, there is evidence of mean reversion in household expenditure in the cases of Benin, Cameroon, Central African Republic, Democratic Republic of Congo, Malawi, Nigeria, Rwanda, Seychelles, the Gambia, Togo, and Uganda, accounting for approximately 28.95% of the 38 countries considered. Also, using the result from the Fourier form of the ADF framework (FADF), we found evidence of mean reversion in the household expenditure of Benin, Cameroon, Central African Republic, Congo, Democratic Republic of Congo, Malawi, Seychelles, Sierra Leone, Togo and Uganda (approximately 26.32% of total cases considered). It was observed that there was only a sparing distinction between the results of the ADF and that of the FADF. The unit root test results by the ADF and the FADF both display evidence of mean reversion in Benin, Cameroon, Central African Republic, Democratic Republic of Congo, Malawi, Seychelles, Togo, and Uganda (approximately 21.1%). Considering the ADF-SB test, it was discovered that there has been an increase in the number of rejections of the unit-roots. More importantly, the affected cases under the FADF test are almost exhaustively a subset of the rejection cases under the ADF-SB test, save Sierra Leone. The number of cases that indicated mean reversion under the ADF-SB test is 24 (approximately 63.16% of the total cases). The evidence of an increase in the number of unit root rejection is because the ADF-SB test accounts for instantaneous breaks, while the FADF test does not. Furthermore, by using the FADF-SB test, the number of mean reversion cases increased to 33 (approximately 86.84% of the total cases), excluding just five countries – Egypt, Ghana, Madagascar, Nigeria, and Tunisia. The increment is also evident from the fact that the FADF-SB test allows for a smooth break. However, as a result of the consistency in the non-rejection of unit roots in the three countries, Egypt, Madagascar, and Tunisia, for the FADF, ADF-SB, and FADF-SB tests, there is, in turn, a nonrejection of the hypothesis of unit root for the household consumption expenditures in three countries (Egypt, Madagascar, and Tunisia), and this indicates non-mean reversion.

**Table 3:** Result of ADF, FADF, ADF-SB, and FADF-SB unit root tests

Country	Code	ADF	FADF	K	ADF-SB	T <sub>B</sub>	λ	FADF-SB	T <sub>B</sub>	λ	K
Algeria	DZA	-1.407	-3.641	1	-3.165	1999	34	<b>-5.702</b>	<b>2008</b>	<b>66</b>	<b>1</b>
Benin	BEN	<b>-3.660</b>	<b>-4.859</b>	<b>1</b>	<b>-4.377</b>	<b>2004</b>	<b>52</b>	<b>-5.660</b>	<b>2015</b>	<b>90</b>	<b>1</b>
Botswana	BWA	-2.344	-3.313	1	-4.860	2008	66	<b>-5.133</b>	<b>2008</b>	<b>66</b>	<b>1</b>
Burkina Faso	BFA	-1.937	-3.129	1	-3.760	2009	69	<b>-4.659</b>	<b>1993</b>	<b>14</b>	<b>1</b>
Burundi	BDI	-3.243	-4.413	1	<b>-4.381</b>	<b>2014</b>	<b>86</b>	<b>-5.049</b>	<b>2012</b>	<b>79</b>	<b>1</b>
Cameroon	CMR	<b>-4.123</b>	<b>-5.374</b>	<b>1</b>	<b>-7.264</b>	<b>1995</b>	<b>21</b>	<b>-8.236</b>	<b>1995</b>	<b>21</b>	<b>2</b>
Central Afri. R.	CAF	<b>-3.851</b>	<b>-4.617</b>	<b>2</b>	<b>-4.186</b>	<b>2008</b>	<b>66</b>	<b>-5.958</b>	<b>2009</b>	<b>69</b>	<b>1</b>
Chad	TCD	-2.803	-3.840	2	<b>-9.969</b>	<b>2002</b>	<b>45</b>	<b>-8.974</b>	<b>2002</b>	<b>45</b>	<b>1</b>
Congo	COG	-3.409	<b>-4.636</b>	<b>1</b>	<b>-4.905</b>	<b>1998</b>	<b>31</b>	<b>-10.015</b>	<b>2014</b>	<b>86</b>	<b>2</b>
Congo, Dem. Rep.	COD	<b>-3.762</b>	<b>-6.406</b>	<b>1</b>	<b>-5.799</b>	<b>1996</b>	<b>24</b>	<b>-7.785</b>	<b>1995</b>	<b>21</b>	<b>1</b>
Egypt	EGY	-1.056	-0.311	2	-3.429	2011	76	-3.654	1998	31	2
Gabon	GAB	-2.588	-3.952	1	<b>-4.458</b>	<b>2014</b>	<b>86</b>	<b>-5.059</b>	<b>2008</b>	<b>66</b>	<b>2</b>
Ghana	GHA	-1.469	-2.841	1	<b>-4.565</b>	<b>2012</b>	<b>79</b>	-4.168	2012	79	2
Guinea	GIN	-2.407	-3.253	2	<b>-4.060</b>	<b>2005</b>	<b>55</b>	<b>-4.474</b>	<b>2006</b>	<b>59</b>	<b>2</b>
Guinea-Bissau	GNB	-2.221	-4.044	1	<b>-4.245</b>	<b>1997</b>	<b>28</b>	<b>-6.072</b>	<b>2001</b>	<b>41</b>	<b>1</b>
Kenya	KEN	-2.232	-3.386	1	<b>-4.835</b>	<b>1994</b>	<b>17</b>	<b>-5.749</b>	<b>1995</b>	<b>21</b>	<b>2</b>
Madagascar	MDG	-2.704	-2.973	1	-3.788	2011	76	-4.084	2006	59	1
Malawi	MWI	<b>-4.288</b>	<b>-4.592</b>	<b>2</b>	<b>-5.261</b>	<b>2006</b>	<b>59</b>	<b>-5.092</b>	<b>2012</b>	<b>79</b>	<b>2</b>
Mali	MLI	-1.793	-3.714	1	-3.599	2012	79	<b>-5.283</b>	<b>2012</b>	<b>79</b>	<b>1</b>
Mauritania	MRT	-2.347	-3.597	1	<b>-3.709</b>	<b>1994</b>	<b>17</b>	<b>-7.515</b>	<b>1993</b>	<b>14</b>	<b>1</b>
Mauritius	MUS	-2.249	-4.326	1	<b>-4.223</b>	<b>2007</b>	<b>62</b>	<b>-6.289</b>	<b>2000</b>	<b>38</b>	<b>1</b>
Morocco	MAR	-1.881	-2.330	1	-3.485	2010	72	<b>-4.737</b>	<b>2011</b>	<b>76</b>	<b>2</b>
Mozambique	MOZ	-2.116	-3.483	1	<b>-3.952</b>	<b>1994</b>	<b>17</b>	<b>-5.645</b>	<b>2001</b>	<b>41</b>	<b>1</b>
Namibia	NAM	-2.362	-2.588	2	-3.691	2003	48	<b>-5.829</b>	<b>2008</b>	<b>66</b>	<b>2</b>
Niger	NER	-2.735	-3.680	2	<b>-4.093</b>	<b>2011</b>	<b>76</b>	<b>-5.238</b>	<b>2011</b>	<b>76</b>	<b>2</b>
Nigeria	NGA	<b>-4.036</b>	-3.789	1	<b>-4.797</b>	<b>1998</b>	<b>31</b>	-4.623	1999	34	1
Rwanda	RWA	<b>-3.987</b>	-3.943	1	<b>-8.898</b>	<b>1994</b>	<b>17</b>	<b>-11.703</b>	<b>1994</b>	<b>17</b>	<b>2</b>
Senegal	SEN	-1.621	-4.206	1	-3.062	2005	55	<b>-4.798</b>	<b>2006</b>	<b>59</b>	<b>1</b>
Seychelles	SYC	<b>-3.957</b>	<b>-4.449</b>	<b>2</b>	<b>-5.132</b>	<b>2004</b>	<b>52</b>	<b>-5.749</b>	<b>2017</b>	<b>97</b>	<b>1</b>
Sierra Leone	SLE	-2.339	<b>-4.626</b>	<b>2</b>	-3.389	2003	48	<b>-6.627</b>	<b>2009</b>	<b>69</b>	<b>2</b>
South Africa	ZAF	-2.799	-3.935	1	<b>-4.259</b>	<b>2001</b>	<b>41</b>	<b>-5.116</b>	<b>2011</b>	<b>76</b>	<b>1</b>
Sudan	SDN	-2.885	-3.668	1	<b>-5.116</b>	<b>1999</b>	<b>34</b>	<b>-6.787</b>	<b>1999</b>	<b>34</b>	<b>1</b>
Tanzania	TZA	-0.363	-4.019	1	-2.110	1997	28	<b>-4.793</b>	<b>2014</b>	<b>86</b>	<b>1</b>
The Gambia	GMB	<b>-3.512</b>	-3.502	1	-3.826	1996	24	<b>-5.459</b>	<b>2011</b>	<b>76</b>	<b>1</b>
Togo	TGO	<b>-3.969</b>	<b>-4.571</b>	<b>2</b>	<b>-5.034</b>	<b>1999</b>	<b>34</b>	<b>-5.237</b>	<b>1999</b>	<b>34</b>	<b>2</b>
Tunisia	TUN	-0.968	-2.755	1	-3.490	2010	72	-4.078	2006	59	1
Uganda	UGA	<b>-3.733</b>	<b>-5.037</b>	<b>2</b>	<b>-4.344</b>	<b>1999</b>	<b>34</b>	<b>-5.395</b>	<b>1999</b>	<b>34</b>	<b>2</b>
Zimbabwe	ZWE	-0.794	-4.397	1	-2.496	2001	41	<b>-5.367</b>	<b>2009</b>	<b>69</b>	<b>1</b>

**Note:** Bolded figures denote that the test is significant at 5% level

Afterwards, the F-test statistic is used to juxtapose the different pairs of unrestricted and restricted model constructs to determine which of the unit root tests considered in the analysis would yield the most viable and reliable mean reversion decision, and its consistency in doing so compared to other tests. This is to determine the test that would

caption the sum of squares regression variation in the household consumption expenditure excellently. Based on the result presented in Table 4, it was discovered that the F-test ( $F_{FADF\_ADF}$ ), which investigates the improvement in the FADF over the ADF test, shows significant improvement in FADF in just 3 of the 38 cases, which are Algeria, Sierra Leone and Zimbabwe. This apparently indicates a high power of the Classical ADF test over its Fourier form (FADF) test. Considering the F-test ( $F_{ADF-SB\_ADF}$ ), the test investigates the significant improvement in the ADF-SB over the ADF. We discovered that there has been an appreciable improvement with respect to 18 cases over the ADF. Considering all the results, it is evident that the FADF-SB test performed highly well over the other three tests, ADF, FADF, and ADF-SB, in all the African countries examined in the study, except for Benin, Burundi, Nigeria, Seychelles, South Africa, Tanzania and Togo (7). Hence, it is safe to say that the FADF-SB unit root test is more reliable and preferable compared to others as displayed in Table 4. Additionally, it is evidently shown, in the result, that the mean reversion hypothesis is significantly affected by the availability of structural breaks. This thereby improve the power of the Fourier function test when combined with structural breaks in unit root testing framework.

**Table 4:** F tests

Country	Code	$F_{FADF\_ADF}$	$F_{ADF-SB\_ADF}$	$F_{FADF-SB\_ADF}$	$F_{FADF-SB\_FADF}$	$F_{FADF-SB\_ADF-SB}$
Algeria	DZA	<b>12.278</b>	7.117	<b>13.745</b>	7.873	<b>27.359</b>
Benin	BEN	4.889	2.918	5.699	4.866	5.524
Botswana	BWA	4.102	<b>12.516</b>	<b>7.448</b>	<b>8.219</b>	<b>14.606</b>
Burkina Faso	BFA	4.114	5.281	5.363	<b>5.134</b>	<b>10.265</b>
Burundi	BDI	3.783	4.650	3.238	2.274	5.284
Cameroon	CMR	4.672	<b>14.623</b>	<b>10.268</b>	<b>11.570</b>	<b>20.464</b>
Central Afr.	CAF	2.970	1.952	5.428	<b>6.471</b>	<b>8.072</b>
Chad	TCO	3.669	<b>124.785</b>	<b>85.869</b>	<b>127.662</b>	<b>160.603</b>
Congo	COG	5.342	5.188	<b>17.505</b>	<b>20.575</b>	<b>21.745</b>
Congo, Dem.	COD	9.624	<b>10.251</b>	<b>9.084</b>	5.107	<b>15.383</b>
Egypt	EGY	5.401	<b>7.996</b>	<b>10.511</b>	<b>10.948</b>	<b>11.759</b>
Gabon	GAB	4.265	<b>6.014</b>	5.039	4.511	<b>9.563</b>
Ghana	GHA	4.087	<b>12.730</b>	<b>6.197</b>	6.392	<b>11.836</b>
Guinea	GIN	3.067	<b>9.599</b>	<b>8.129</b>	<b>10.625</b>	<b>9.893</b>
Guinea-Bissau	GNB	5.553	<b>7.701</b>	<b>8.345</b>	<b>7.836</b>	<b>15.669</b>
Kenya	KEN	3.870	<b>11.379</b>	<b>12.116</b>	<b>15.487</b>	<b>22.743</b>
Madagascar	MDG	1.048	3.352	4.878	<b>8.063</b>	<b>9.296</b>
Malawi	MWI	1.398	3.700	3.402	4.928	<b>6.371</b>
Mali	MLI	7.765	4.892	<b>9.924</b>	<b>7.616</b>	<b>18.965</b>
Mauritania	MRT	4.730	<b>20.134</b>	<b>13.989</b>	<b>16.764</b>	<b>26.844</b>
Mauritius	MUS	7.520	<b>7.592</b>	<b>10.179</b>	<b>8.158</b>	<b>20.335</b>
Morocco	MAR	4.238	4.346	<b>7.406</b>	<b>7.995</b>	<b>14.594</b>
Mozambique	MOZ	5.443	<b>7.579</b>	<b>10.276</b>	<b>10.577</b>	<b>19.916</b>
Namibia	NAM	2.178	4.684	<b>8.849</b>	<b>13.208</b>	<b>16.615</b>
Niger	NER	4.103	4.692	<b>5.746</b>	5.710	<b>11.234</b>
Nigeria	NGA	0.408	2.910	1.966	3.437	3.726

**Table 4:** F tests (cont.)

Country	Code	F <sub>FADF_ADF</sub>	F <sub>ADF-SB_ADF</sub>	F <sub>FADF-SB_ADF</sub>	F <sub>FADF-SB_FADF</sub>	F <sub>FADF-SB_ADF-SB</sub>
Rwanda	RWA	0.996	<b>180.571</b>	<b>161.930</b>	<b>297.207</b>	<b>320.899</b>
Senegal	SEN	9.857	3.724	<b>6.945</b>	2.634	<b>13.667</b>
Seychelles	SYC	2.233	4.596	4.032	5.045	5.952
Sierra Leone	SLE	<b>10.677</b>	3.193	<b>11.542</b>	<b>6.915</b>	<b>21.331</b>
South Africa	ZAF	4.526	4.888	4.887	4.048	7.238
Sudan	SDN	2.999	<b>9.412</b>	<b>9.951</b>	<b>13.614</b>	<b>19.163</b>
Tanzania	TZA	9.828	3.223	<b>7.342</b>	3.079	-1.169
Gambia	GMB	0.692	2.047	5.058	<b>8.946</b>	<b>8.930</b>
Togo	TGO	2.488	3.964	2.725	2.613	5.417
Tunisia	TUN	4.212	<b>10.498</b>	4.912	4.375	<b>7.984</b>
Uganda	UGA	5.559	2.137	3.683	1.544	<b>7.157</b>
Zimbabwe	ZWE	<b>14.221</b>	<b>6.210</b>	<b>11.438</b>	4.423	<b>16.696</b>

**Note:** In bold indicates significance at 5% level. See Furuoka (2017a) for critical values.

#### 4. Conclusion

The study examines the evidence of mean reversion or non-mean reversion in household expenditures in selected thirty-eight (38) countries across Africa between 1990 and 2018 using the Fourier unit root test with breaks (FADF-SB). The test procedure works quite well in the presence of a small sample size and it is capable of controlling for smooth breaks based on the Fourier function in the test regression. Other unit root tests, the ADF, FADF, and ADF-SB, are restricted versions to the FADF-SB, which further gives the general test appealing properties. An F test that determines the superiority of FADF-SB and ADF-SB is also presented. On applying the traditional ADF unit root test, the authors discovered that only 16 of the considered cases indicate significance. That is, about 42% signify evidence of mean reversion in the time series of household expenditures across African countries considered. The three tests, FADF, ADF-SB, and FADF-SB rejected unit root hypotheses in 26.31%, 63.16%, and 86.84% of the total cases, respectively. Based on these results, the household expenditure rate in most of the African countries is mean-reverting for the period considered in this study. The FADF-SB test outperformed others in the majority of the African countries considered. Based on the results of only the FADF-SB test, the non-mean reversion hypothesis holds in five (5) of the thirty-eight (38) African countries examined in this study. Meanwhile, based on the decisions of nonrejection of unit root by the four tests considered, non-mean reversion exists in only three countries, that is, in Egypt, Madagascar, and Tunisia. In these cases of non-mean reversion, the household expenditure rates do not revert to their mean levels, which is the likelihood that these countries may experience the persistence of shocks for a longer period. The implication herewith requires strong public policy actions to address the

household expenditure shock. Depending on the nature of the shock, a strong check and balance are needed to be put in place. In cases where the household expenditure rate is lower compared to the Gross Domestic Product (GDP) per capita, the shocks could mean evidence of economic development. However, in cases where the household expenditure rate is higher, the shock could mean the evidence of GDP per capita debt. Considering the robustness investigation to ascertain the most powerful and preferable test, the result indicates the dominance of the FADF-SB over other tests.

## References

- African Development Bank, (2012). A Comparison of Real Household Consumption Expenditures and Price Levels in Africa. Tunis: African Development Bank.
- Akhand, A., (2011). Household Consumption Behaviour in Indonesia, Newcastle Business School, The University of Newcastle The University Drive, Callaghan, NSW 2308, Australia.
- Awolaja, O. G., Yaya, O. S., Vo, X. V., Ogbonna, A. E. and JOSEPH, S. O., (2021). Unemployment Hysteresis in Middle East and North Africa Countries: Panel SUR-based Unit root test with a Fourier function. *Middle East Development Journal*, pp. 13 (2), 318–334.
- Becker, R., Enders, W., and Lee, J., (2006). A stationarity test in the presence of an unknown number of smooth breaks. *Journal of Time Series Analysis*, 27(3), pp. 381–401.
- Ben, E., (2015). A test of the household income process using consumption and wealth data, *European Economic Review*, <http://dx.doi.org/10.1016/j.eurocorev.2015.05.003>.
- Box, G. E. P., Jenkins, G. M., Reinsel, G. C., and Ljung, C. M., (2015). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, UK.
- Enders, W., Lee, J., (2012a). A unit root test using a Fourier series to approximate smooth breaks. *Oxford Bulletin of Economics and Statistics*, 74, pp. 574–599.
- Enders, W., Lee, J., (2012b). The flexible Fourier form and Dickey-Fuller-type unit root tests. *Economic Letters*, 117, pp. 196–199.
- Furuoka, F., (2017a). A new approach to testing unemployment hysteresis. *Empirical Economics*, 53(3), pp. 1253–1280.
- Furuoka, F., (2017b). Unemployment Dynamics in the Asia-Pacific region: A Preliminary investigation. *The Singapore Economic Review*, 62(5), pp. 983–1016.



- Fumitaka, F., Pui, K. L., Ezeoke, C., Jacob, R. I. and Yaya, O. S., (2020). Growth slowdowns and the middle-income trap: Evidence based on new Unit root framework. *The Singapore Economic Review*. <https://doi.org/10.1142/S0217590820500083>.
- Gil-Alana, L. A., Yaya, O. S., (2021). Testing Fractional Unit Roots with Non-linear Smooth Break Approximations using Fourier functions. *Journal of Applied Statistics*, 48 (13-15), pp. 2542–2559
- Giray, G., (2013). Stochastic properties of the consumption-income ratios in Central and Eastern European Countries *Zb. rad. Ekon. fak. Rij.*, vol. 31, sv. 2, pp. 193–207
- Lozano-Gracia, N., Young, C., (2014). Housing Consumption and Urbanization. *Policy Research Working Paper.*, 7712, pp. 4–5.
- Mahdavi, D. B., (2013). *The Non-Misleading Value of Inferred Correlation: An Introduction to the Cointelation Model*. Wilmott., 1, pp. 50–61.
- Nik, M., Nik, H.R., Noorhaslinda, A. and Nasir, A. (2011). Household Income and Expenditure Relationships: A Simultaneous Equation Approach. *Chinese Business Review*, ISSN 1537-1506. pp. 395–405.
- OECD, (2020). *Household Spending (indicator)*. DOI: 10.1787/b5f46047,
- Thankgod, O. (2014). *Private Consumption Expenditure Function in Nigeria: Evidence from the Keynes' Absolute Income Hypothesis*.
- Yaya, O. S., Furuoka, F., Ling, P. K., Jacob, R. I. and Ezeoke, C. M. R. (2020). Investigating Asian Regional Income Convergence using Fourier Unit Root test with Break. *International Economics*, 161, pp. 120–129.
- Yaya, O. S., Ogbonna, A. E. and Mudida, R. (2019a). Hysteresis of Unemployment rate in Africa: New Findings from Fourier ADF test. *Quality and Quantity International Journal of Methodology*, 53(6), pp. 2781–2795.
- Yaya, O. S., Ling, P. K., Furuoka, F., Ezeoke, C. M. R. and Jacob, R. I. (2019b). Can Western African countries catch up with Nigeria? Evidence from Smooth Nonlinearity method in Fractional Unit root framework. *International Economics*, 158, pp. 51–63.
- Yaya, O. S. (2018). Another Look at the Stationarity of Inflation rates in OECD countries: Application of Structural break-GARCH-based unit root tests. *Statistics in Transition new series*, 19(3), pp. 477–492

- Yaya, O. S., Otekunrin, O. A. and Ogbonna, A. E. (2021). Life Expectancy in West African countries: Evidence of Convergence and Catching up with the North. *Statistics in Transition*, 22(1), pp. 75–88.
- Zivot, E., Andrews, D. W. K. (1992). Further evidence on Great Crash, the oil price shock, and the unit root hypothesis. *Journal of Business and Economic Statistics*, 10, pp. 251–270.

## Does economic freedom promote financial development? Evidence from EU countries

Anand Sharma<sup>1</sup>, Vipin Sharma<sup>2</sup>, Shekhar Tokas<sup>3</sup>

### Abstract

This study empirically investigates the relationship between economic freedom and financial development in EU countries. Using panel data covering the years 2000–2017 and employing fixed effects, random effects, and the generalised method of moments (GMM), the paper examines the effect of economic freedom on financial development. The research results demonstrate that greater economic freedom is conducive to financial development in the EU. These findings remain robust to the use of an alternative index of economic freedom. The results imply that policies which promote economic freedom are likely to raise the level of a country's financial development.

**Key words:** economic freedom, financial development, panel data, EU.

### 1. Introduction

The focus on financial development has increased considerably in the recent decades. Evidence suggests that financial development has a favourable effect on economic growth, poverty, and inequality (e.g. Levine, 1997; Rajan and Zingales, 1998; Kappel, 2010; Guru and Yadav, 2019). Financial development also encourages the growth of small and medium enterprises, and it is an important component of economic development (World Bank, 2016). Thus, enhancing the strength of financial markets and institutions becomes imperative for achieving higher growth and development in a country. Financial development is closely linked to the quality of economic institutions of a country. Economic freedom is an indicator of this institutional quality (Hall et al., 2019; Sharma, 2020) and it measures the extent to which

---

<sup>1</sup> O. P. Jindal Global University, Sonipat, India. E-mail: [anandsharma@jgu.edu.in](mailto:anandsharma@jgu.edu.in).  
ORCID: <https://orcid.org/0009-0007-7467-6918>.

<sup>2</sup> University School of Business, Chandigarh University, Mohali, India. E-mail: [vipin.e9155@cumail.in](mailto:vipin.e9155@cumail.in).  
ORCID: <https://orcid.org/0000-0002-4215-9808>.

<sup>3</sup> Dr. B. R. Ambedkar University, Delhi, India. E-mail: [shekhar@aud.ac.in](mailto:shekhar@aud.ac.in).  
ORCID: <https://orcid.org/0000-0001-8470-3392>.



the institutions and policies of a country are market-oriented (Stroup, 2007; Angulo-Guerrero et al., 2017). Economic freedom includes five key areas: the size of government, legal system and property rights, sound money, free trade, and regulation (Gwartney et al. 2021).

In theory, economic freedom may affect financial development due to various reasons. Well-defined property rights are an important feature of financial transactions. A robust legal system and property rights help in dealing with asymmetric information and thus lower the costs associated with financial transactions (Fernández and Tamayo, 2017). The existence of these institutions raises the availability and efficacy of external finance and affects the extent of appropriation (Fergusson, 2006; Beck and Levine, 2005). A strong legal framework and property rights mechanism also aid in the enforcement of financial contracts and increase the confidence of different stakeholders in the financial sector. Economic freedom in the form of free trade promotes competition and restricts the rent-seeking behaviour of incumbent elites and thus enhances financial development (Law, 2008). This argument is based on the 'interest group' theory advanced by Rajan and Zingales (2003). Sometimes, free trade is associated with an increased risk arising out of fluctuations in the global economy and therefore, it may lead to the development of a financial sector in the economy to combat these risks (Svaleryd and Vlachos, 2002). Trade openness also exerts a positive effect on financial depth (Huang and Temple, 2005) and thus promotes financial development. Economic freedom resulting from sound money may promote financial development as a stable and low rate of inflation raises the real return on assets and avoids the adverse selection problems (Fernández and Tamayo, 2017; Feldstein, 1980). In this paper, we examine if greater economic freedom is associated with a higher level of financial development in EU countries.

Most of the empirical studies have linked economic freedom with growth (Gwartney et al., 1999; Bergh and Bjørnskov, 2021), entrepreneurship (Nyström, 2008; Sweidan, 2021), corruption (Graeff and Mehlkop, 2003; Thach and Ngoc, 2021), education (Feldmann, 2017) and health (Stroup, 2007; Sharma, 2020), among other variables. However, the literature which analyses the relationship between economic freedom and financial development is quite scarce (e.g. Enowbi-Batuo and Kupukile, 2010; Hafer, 2013; Khan et al., 2021). Hafer (2013) investigated the effect of economic freedom on financial development for 81 countries from 1980 to 2009. He found that the initial level of economic freedom has a positive effect on subsequent financial development. Several scholars have examined this connection in the context of developing and underdeveloped countries. For example, Khan et al. (2021) studied the effect of economic freedom on financial development for 87 developing countries from 1984 to 2018. They used the panel threshold model and found that economic freedom exerts a favourable effect on financial development. Enowbi-Batuo and Kupukile (2010)

examined the interaction between economic freedom, political freedom and financial development for the African countries from 1990 to 2005. They utilized difference-in-difference and panel regression methods and showed that economic freedom enhanced financial development in these countries.

The empirical literature has also analysed the effect of economic freedom on banking crises (e.g. Baier et al., 2012; Shehzad and de Haan, 2009) and economic crises (e.g. Bjørnskov, 2016; Giannone et al., 2011). Most of these studies found a favourable effect of economic freedom on crises. For example, Baier et al. (2012) analysed the data on banking crises from 1976 to 2008 and observed that economic freedom significantly lowers the likelihood of a banking crisis. Shehzad and de Haan (2009) examined the relationship between economic freedom and crises for the developed and developing countries from 1973 to 2002 and found a similar effect. Bjørnskov (2016) studied the association between economic freedom and economic crises for 175 countries from 1993 to 2010. He found a weak relationship between economic freedom and the occurrence of an economic crisis but concluded that freer countries faced smaller crises and a quicker recovery. Studies have also established a positive effect of economic freedom on credit allocation (e.g. Crabb 2008; Hartarska and Nadolnyak 2007) and bond ratings (e.g. Belasen et al., 2015; Dove, 2017).

Very few studies have examined the relationship between economic freedom and financial development for the developed countries and to the best of our knowledge, no study has analysed this linkage in the context of EU countries. Further, most of the studies have not used alternative measures of economic freedom to assess the robustness of the results. The main objective of this study is to analyse the effect of economic freedom on financial development for the EU countries from 2000 to 2017. This paper fills several important research gaps in the financial economics literature. First, this paper uses alternative measures of economic freedom to analyse its effect on financial development. Second, this paper focuses on the developed EU countries. Third, this study addresses the endogeneity concerns by using the GMM method.

The remaining paper is organised as follows. Section 2 explains the variables and lists the data sources. Section 3 describes the methodology used in the paper. Section 4 presents the results and discusses the key policy implications. The last section presents the concluding remarks.

## **2. Data**

This paper uses the data on economic freedom, financial development, and the relevant control variables for the 27 EU countries from 2000 to 2017. The data on economic freedom are published by two institutes viz. the Fraser Institute and the Heritage Foundation. We primarily rely on the economic freedom index released by

the Fraser Institute due to its robustness and acceptance in the literature (e.g. Easton and Walker, 1992; Angulo-Guerrero et al., 2017). This index measures the degree to which individuals are protected from expropriation and can make their economic decisions freely. This index takes the values between 0 and 10 with higher values representing greater economic freedom. It has five areas and consists of 44 variables (Fraser Institute, 2022). We also employ the Heritage Foundation index of economic freedom. This index is widely used in the literature (e.g. Crabb 2008; Bjørnskov, 2016). This index comprises 12 areas and it ranges from 0 to 100 with higher values implying greater economic freedom (Heritage Foundation, 2022).

We obtain the data on the dependent variable viz. financial development index from the International Monetary Fund (IMF, 2022). The literature views this index as detailed and multidimensional (e.g. Khan et al., 2021; Svirydzenka 2016). This index ranges from 0 to 1 with higher values denoting greater financial development. The data on per capita GDP, foreign direct investment (FDI), and consumer price index (CPI) are taken from the World Development indicators of the World Bank. The net interest margin data is obtained from the IMF and the democracy (political rights) index is collected from the Freedom House (2022). We recode the political rights index so that larger values show the presence of a greater democratic environment. The recoded index takes the values from 1 to 7. Table 1 presents the descriptive statistics for the variables used in this paper.

**Table 1:** Descriptive Statistics

Variable	Obs	Mean	S.D.	Min	Max
FD	476	.561	.198	.13	.91
EF (Fraser)	476	7.575	.4	5.55	8.32
EF (Heritage)	476	67.599	6.521	47.3	82.6
Per capita GDP	476	37685.12	17680.28	10201.28	115000
FDI	476	13.316	40.693	-58.323	449.083
NIM	476	2.371	1.437	.126	9.908
Democracy	476	6.866	.36	5	7
CPI	476	95.06	13.319	31.982	115.455

### 3. Methods

We rely on the existing literature and specify the following empirical model to determine the effect of economic freedom on financial development (e.g. Enowbi-Batuo and Kupukile, 2010; Hafer, 2013; Khan et al., 2021)

$$Y_{it} = \beta_0 + \beta_1 \text{Economic Freedom}_{it} + \beta_2 Z_{it} + \gamma_i + \varepsilon_{it} \quad (1)$$

Where  $Y_{it}$  is the overall index of financial development and  $\text{Economic Freedom}_{it}$  is the index of economic freedom in country 'i' at year 't'.  $Z_{it}$  shows the standard control variables and includes per capita GDP, FDI, net interest margin, democracy, and CPI. The definitions of these variables are provided in Appendix Table 1.  $\gamma_i$  represent the fixed effects and  $\varepsilon_{it}$  denotes the error term. The use of pooled OLS method produces biased and inconsistent estimates due to heterogeneity bias (Wooldridge, 2009). Thus, we use fixed effects and random effects models to deal with the unobserved heterogeneity. We select the appropriate model using the Hausman test. Additionally, we formulate the following dynamic panel data (DPD) model containing the lagged financial development index,  $Y_{it-1}$  as one of the explanatory variables.

$$Y_{it} = \beta_0 + \rho Y_{it-1} + \beta_1 \text{Economic Freedom}_{it} + \beta_2 Z_{it} + \gamma_i + \varepsilon_{it} \quad (2)$$

Using fixed effects and random effects methods to estimate this model is problematic due to the correlation between the lagged dependent variable and fixed effects in the error term. We resolve this endogeneity by transforming the original equation by taking the first differences (Roodman, 2009; Greene, 2003). There are no fixed effects in the transformed equation and lagged levels are taken as instruments of the first-differenced variables (Baum, 2013). This generalized method of moments approach is based on the seminal work of Arellano and Bond (1991) and takes the following form in this case:

$$\Delta Y_{it} = \rho \Delta Y_{it-1} + \beta_1 \Delta \text{Economic Freedom}_{it} + \beta_2 \Delta Z_{it} + \varepsilon_{it} \quad (3)$$

We follow the approach outlined by Roodman (2009) to implement a two-step difference GMM and prefer it over one-step GMM as the former is robust to heteroskedasticity and autocorrelation. Arellano and Bover (1995) and Blundell and Bond (1998) developed a system GMM approach that includes lagged differences as instruments in addition to the lagged levels. However, in this case, we choose difference GMM over system GMM as the latter uses a larger number of instruments. The overidentifying restrictions and the instruments may not remain valid when the number of instruments exceeds the number of groups (Bondarenko, 2012). We conduct the Hansen test of overidentifying restrictions to determine the validity of instruments. We also carry out the Arellano and Bond test to detect the presence of serial correlation of second-order in residuals.

#### 4. Results and discussion

In this section, we present the results of the empirical model. All the regression results report the standardized coefficients. Table 2 reports the fixed effects and random effects results. The Hausman test supports the use of the FE model as the p-value is less than 0.05. The FE results in column (1) show that economic freedom has a positive

impact on the financial development in the EU countries. The coefficient on economic freedom is 0.061 and it is significant at a 5% level. This coefficient implies that one standard deviation improvement in economic freedom is associated with a 0.061 standard deviation increase in financial development. The RE results in column (2) also support this finding. Most of the control variables turn out to be significant in both models. For example, an increase in net interest margin is associated with a decline in financial development. Column (1) shows that the coefficient on net interest margin is -0.11 and it is significant at 1% level. This implies that one standard deviation increase in net interest margin is associated with a 0.11 standard deviation decline in financial development. A stronger democracy is associated with an improvement in financial development. An increase in per capita GDP also improves the financial development in the EU countries. However, this coefficient only turns out to be significant in the RE model. The remaining two control variables viz. FDI and CPI are insignificant in both models.

**Table 2:** Economic Freedom (Fraser) and Financial Development: Fixed and Random effects

Specification	(1) FE	(2) RE
Economic freedom (Fraser Institute)	0.061** (0.025)	0.051** (0.026)
Per capita GDP	0.184 (0.135)	0.277** (0.112)
Foreign Direct Investment	-0.001 (0.011)	0.002 (0.011)
Net Interest Margin	-0.110*** (0.031)	-0.130*** (0.029)
Democracy	0.046** (0.019)	0.049*** (0.019)
Consumer Price Index	0.008 (0.031)	-0.007 (0.028)
Obs.	476	476
Adjusted R <sup>2</sup>	0.26	-

Robust standard errors are in parenthesis

\*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

We examine the robustness of the results presented in Table 2 in two ways. First, we use Heritage Foundation's index of economic freedom. Second, we employ the two-step difference GMM method to tackle the endogeneity concerns. Table 3 reports both the FE and RE results with the index of economic freedom prepared by the Heritage Foundation. These results corroborate the findings obtained earlier. We find a positive



effect on financial development by using the alternative index of economic freedom. As column (1) shows one standard deviation increase in the economic freedom index is associated with a 0.106 standard deviation rise in financial development. The coefficients on control variables are broadly similar. A better democratic environment and an increase in per capita GDP leads to greater financial development, whereas a rise in net interest margin retards it. We focus on the FE results as the Hausman test advocates the use of the FE model.

**Table 3:** Economic Freedom (Heritage) and Financial Development: Fixed and Random effects

Specification	(1) FE	(2) RE
Economic freedom (Heritage)	0.106** (0.044)	0.087* (0.045)
Per capita GDP	0.181 (0.135)	0.273** (0.115)
Foreign Direct Investment	0.002 (0.007)	0.004 (0.008)
Net Interest Margin	-0.101*** (0.029)	-0.122*** (0.027)
Democracy	0.043* (0.021)	0.046** (0.021)
Consumer Price Index	0.007 (0.030)	-0.007 (0.028)
Obs.	476	476
Adjusted R <sup>2</sup>	0.27	-

Robust standard errors are in parenthesis

\*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

Table 4 presents the estimation of the empirical model using the two-step difference GMM. These results confirm the robustness of our previous findings presented in Table 2. The number of instruments is 23 and the number of groups (countries) is 27. The Arellano-Bond test for second-order serial correlation exhibits a p-value of 0.11 and thus indicates the absence of second-order serial correlation. The Hansen test of overidentifying restrictions shows a p-value of 0.365 and indicates that the instruments are valid. The GMM results also highlight that economic freedom has a positive effect on financial development. The coefficient on the economic freedom index is 0.071 and it shows that one standard deviation increase in economic freedom is associated with 0.071 standard deviation improvement in financial development. The coefficient on lagged financial development index is positive and significant. Other control variables turn out to be insignificant in this model.

Our findings validate the results of previous studies, which have also found a positive effect of economic freedom on financial development (e.g. Hafer, 2013; Khan et al., 2021). Our results offer useful policy prescriptions as we focus on a homogenous set of developed EU countries that can better coordinate their policies as compared to other countries. The findings suggest that policies which improve the quality of economic institutions need to be emphasized for enhancing a country's financial development. These policies may take the form of crafting a strong and efficient legal framework, creating a stable macroeconomic environment, reducing unnecessary regulations, and implementing a well-functioning system of property rights. The results also suggest the role of a robust democracy and higher per capita income in improving a country's financial development.

The findings of this paper support the prevalent understanding about the EU countries. The EU countries rank very high in terms of economic freedom and thus have a sound quality of economic institutions. The EU countries are among the economically freest countries of the world and have a strong rule of law, an efficient system of property rights, a stable macroeconomic environment and less regulations. The better quality of economic institutions in the EU countries ensures lower cost of financial transactions and also increases the confidence of various stakeholders in the financial system. The trade openness of the EU countries also has a favourable effect on the financial depth. The extent of financial development in the EU countries is also considerably high and therefore, the positive influence of economic freedom on financial development gets further strengthened in this environment.

## 5. Conclusion

In this study, we explore the impact of economic freedom on financial development for the EU countries from 2000 to 2017. We find that greater economic freedom is associated with an improvement in the financial development in the EU countries. These results suggest significant financial development can be achieved by improving the quality of economic institutions. Therefore, the policymakers should focus on the policies to enhance the level of economic freedom. Our findings remain robust to the use of an alternative index of economic freedom and different techniques viz. fixed effects and GMM.

This paper uses an index of overall financial development as a dependent variable and does not focus on the financial markets and financial institutions sub-indices. Future research may attempt to consider the effect of economic freedom on the development of financial markets and financial institutions. Additionally, scholars may examine the relationship between financial development and the areas of economic freedom to understand their relative importance. This may help the policymakers to

focus on the specific elements of economic freedom, which are most helpful in improving the financial development of a country.

**Table 4:** Two-step difference GMM estimation results: Dep variable: Financial Development

Specification	(1) FD
Lagged Financial Development	0.268** (0.118)
Economic freedom	0.071** (0.029)
Per capita GDP	0.083 (0.143)
Foreign Direct Investment	0.015 (0.017)
Net Interest Margin	-0.016 (0.025)
Democracy	0.022 (0.042)
Consumer Price Index	0.061 (0.036)
Observations	446
No. of instruments	23
No. of Groups	27

Arellano-Bond test for AR(2) in first differences:  $z = -1.59$   $Pr > z = 0.111$

Hansen test of overid. restrictions:  $\chi^2(16) = 17.32$   $Prob > \chi^2 = 0.365$

Robust standard errors are in parenthesis

\*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

## References

- Angulo-Guerrero, M. J., Pérez-Moreno, S., Abad-Guerrero, I. M., (2017). How Economic Freedom affects Opportunity and Necessity Entrepreneurship in the OECD countries. *Journal of Business Research*, 73, pp. 30–37.
- Arellano, M., Bover, O., (1995). Another Look at the Instrumental Variable Estimation of Error-Components Models. *Journal of econometrics*, 68(1) pp. 29–51.
- Arellano, M., Bond, S., (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and An Application to Employment Equations. *The Review of Economic Studies*, 58(2), pp. 277–297.

- Baier, S. L., Clance, M., Dwyer, G. P., (2012). Banking Crises and Economic Freedom. In: Gwartney, J., Lawson, R. A., Hall, J. (Eds.), *Economic Freedom of the World: 2013 Annual Report*, The Fraser Institute, Vancouver, pp. 201–217.
- Baum, C. F., (2013). Dynamic Panel Data Estimators. *Applied Econometrics*, EC823, pp.1–50.
- Beck, T., Levine, R., (2005). *Legal Institutions and Financial Development*. In *Handbook of new institutional economics*, pp. 251–278, Springer.
- Belasen, A. R., Hafer, R. W., Jategaonkar, S. P., (2015). Economic Freedom and State Bond ratings. *Contemporary Economic Policy*, 33(4), pp. 668–677.
- Bergh, A., Bjørnskov, C., (2021). Does Economic Freedom boost Growth for Everyone? *Kyklos*, 74(2), pp. 170–186.
- Bjørnskov, C., (2016). Economic Freedom and Economic Crises. *European Journal of Political Economy*, 45, pp. 11–23.
- Blundell, R., Bond, S., (1998). Initial Conditions and Moment Restrictions in Dynamic Panel Data Models. *Journal of Econometrics*, 87(1), pp. 115–143.
- Bondarenko, E., (2012). *Three Essays on Financial and Trade Integration*. Graduate Theses, Dissertations, and Problem Reports, 150.
- Crabb, P., (2008). Economic Freedom and the Success of Microfinance Institutions. *Journal of Developmental Entrepreneurship*, 13(02), pp. 205–219.
- Dove, J., (2017). The Relationship between Local Government Economic Freedom and Bond Ratings. *Journal of Financial Economic Policy*, 9(4), pp. 435–449.
- Easton, T. S., Walker, A. M. (Eds), (1992). *Rating Global Economic Freedom*. The Fraser Institute, Vancouver, BC.
- Enowbi-Batuo, M., Kupukile, M., (2010). How can Economic and Political Liberalisation improve Financial Development in African Countries? *Journal of Financial Economic Policy*, 2(1), pp. 35–59.
- Feldmann, H., (2017). Economic Freedom and Human Capital Investment. *Journal of Institutional Economics*, 13(2), pp. 421–445.
- Feldstein, M., (1980). Inflation and the Stock Market. *American Economic Review*, 70(5), pp. 839–847
- Fergusson, L., (2006). Institutions for Financial Development: What are they and where do they come from? *Journal of Economic Surveys*, 20(1), pp. 27–70.

- Fernández, A., Tamayo, C. E., (2017). From Institutions to Financial Development and Growth: What are the Links? *Journal of Economic Surveys*, 31(1), pp. 17–57.
- Fraser Institute, (2022). *Economic Freedom of the World*, <https://www.fraserinstitute.org/studies/economic-freedom> (accessed on 20 March 2021).
- Freedom House, (2022). <https://freedomhouse.org/report/freedom-world> (Accessed May 3, 2021)
- Giannone, D., Lenza, M., Reichlin, L., (2011). Market Freedom and the Global Recession. *IMF Economic Review*, 59(1), pp. 111–135.
- Graeff, P., Mehlkop, G., (2003). The Impact of Economic Freedom on Corruption: Different Patterns for Rich and Poor Countries. *European Journal of Political Economy*, 19(3), pp. 605–620.
- Greene, W. H., (2003). *Econometric Analysis*, Pearson Education.
- Guru, B. K., Yadav, I. S., (2019). Financial Development and Economic Growth: Panel Evidence from BRICS. *Journal of Economics, Finance and Administrative Science*, 24(47), pp. 113–126.
- Gwartney, J. D., Lawson, R. A., Holcombe, R. G., (1999). Economic Freedom and the Environment for Economic Growth. *Journal of Institutional and Theoretical Economics*, 155(4), pp. 643–663.
- Gwartney, J. D., Lawson, R. A., Hall, J., Murphy, R., (2021). *Economic Freedom of the World: 2021 Annual Report*, Fraser Institute.
- Hafer, R. W., (2013). Economic Freedom and Financial Development: International evidence. *Cato Journal*, 33, pp. 111.
- Hall, J. C., Lacombe, D. J., Shaughnessy, T. M., (2019). Economic Freedom and Income Levels across US States: A Spatial Panel Data Analysis. *Contemporary Economic Policy*, 37(1), pp. 40–49.
- Hartarska, V., Nadolnyak, D., (2007). Do Regulated Microfinance Institutions achieve Better Sustainability and Outreach? Cross-country Evidence. *Applied Economics*, 39(10), pp. 1207–1222.
- Heritage Foundation, (2022). *Index of Economic Freedom: Methodology*. Retrieved from <https://www.heritage.org/index/about> (accessed 24 March 2021).
- Huang, Y., Temple, J., (2005). Does External Trade Promote Financial Development? *University of Bristol Discussion Paper*, Research-Work in Progress, 575.

- International Monetary Fund, (2022). <https://data.imf.org/?sk=F8032E80-B36C-43B1-AC26-493C5B1CD33B> (Accessed on April 2, 2022)
- Kappel, V., (2010). The effects of financial development on income inequality and poverty. *CER-ETH-Center of Economic Research at ETH Zurich Working Paper*, No. 127.
- Khan, M. A., Islam, M. A., Akbar, U., (2021). Do Economic Freedom matters for Finance in Developing Economies: A Panel Threshold Analysis. *Applied Economics Letters*, 28(10), pp. 840–843.
- Law, S. H., (2008). Does a Country's Openness to Trade and Capital Accounts lead to Financial Development? Evidence from Malaysia. *Asian Economic Journal*, 22(2), pp. 161–177.
- Levine, R., (1997). Financial Development and Economic Growth: Views and Agenda. *Journal of Economic Literature*, 35(2), pp. 688–726.
- Nyström, K., (2008). The Institutions of Economic Freedom and Entrepreneurship: Evidence from Panel Data. *Public choice*, 136(3), pp. 269–282.
- Rajan, R., Zingales, L., (1998). Financial Development and Growth. *American Economic Review*, 88(3), pp. 559–586.
- Rajan, R., Zingales, L., (2003). The Great Reversals: The Politics of Financial Development in the Twentieth Century. *Journal of Financial Economics*, 69(1), pp. 5–50.
- Roodman, D., (2009). How to do Xtabond2: An Introduction to Difference and System GMM in Stata. *The Stata Journal*, 9(1), pp. 86–136.
- Sharma, A., (2020). Does Economic Freedom Improve Health Outcomes in Sub-Saharan Africa? *International Journal of Social Economics*, 47(12), pp. 1633–1649.
- Shehzad, C. T., de Haan, J., (2009). Financial Reform and Banking Crises. *CESifo Working Paper*, No. 2870.
- Stroup, M. D., (2007). Economic Freedom, Democracy, and the Quality of Life. *World Development*, 35(1), pp. 52–66.
- Svaleryd, H., Vlachos, J., (2002). Markets for Risk and Openness to Trade: How are they related? *Journal of International Economics*, 57(2), pp. 369–395.
- Svirydzenka, K., (2016). Introducing a New Broad-based Index of Financial Development. *IMF Working Paper*, 5.

- Sweidan, O. D., (2021). Economic Freedom and Entrepreneurship Rate: Evidence from the US States After the Great Recession. *Journal of the Knowledge Economy*, pp. 1–17.
- Thach, N. N., Ngoc, B. H., (2021). Impact of Economic Freedom on Corruption Revisited in ASEAN Countries: A Bayesian Hierarchical Mixed-Effects Analysis. *Economies*, 9(1), pp. 1–16.
- Wooldridge, J. M., (2009). *Introductory Econometrics: A Modern Approach*. South-Western.
- World Bank, (2016). *Global Financial Development Report: Background*. <https://www.worldbank.org/en/publication/gfdr/gfdr-2016/background/financial-development>
- World Bank, (2022). World Development Indicators, <https://data.worldbank.org/indicator> (accessed 23 March 2021).

## Appendix

**Table A1:** List of variables and their description

Variable	Description	Source
Financial Development (FD)	Index ranging from 0 to 1 with higher values denoting greater financial development	IMF
Economic Freedom (Fraser)	Economic freedom index (ranging from 1 to 10) with higher values denoting greater EF	Fraser Institute
Economic Freedom (Heritage)	Economic freedom index (ranging from 0 to 100) with higher values denoting greater EF	Heritage Foundation
Per capita GDP	Per capita GDP PPP (Constant 2017 international \$)	WDI, World Bank
Foreign Direct Investment	Foreign direct investment, net inflows (% of GDP)	WDI, World Bank
Net Interest Margin	Accounting value of bank's net interest revenue as a share of its average interest-bearing (total earning) assets	IMF
Democracy	Political rights index (ranging from 1 to 7)	Freedom House
Consumer Price Index	Inflation, consumer prices (annual %)	WDI, World Bank



# Bayesian modelling for semi-competing risks data in the presence of censoring

Atanu Bhattacharjee<sup>1</sup>, Rajashree Dey<sup>2</sup>

## Abstract

In biomedical research, challenges to working with multiple events are often observed while dealing with time-to-event data. Studies on prolonged survival duration are prone to having numerous possibilities. In studies on prolonged survival, patients might die of other causes. Sometimes in the survival studies, patients experienced some events (e.g. cancer relapse) before dying within the study period. In this context, the semi-competing risks framework was found useful. Similarly, the prolonged duration of follow-up studies is also affected by censored observation, especially interval censoring, and right censoring. Some conventional approaches work with time-to-event data, like the Cox-proportional hazard model. However, the accelerated failure time (AFT) model is more effective than the Cox model because it overcomes the proportionality hazard assumption. We also observed covariates impacting the time-to-event data measured as the categorical format. No established method currently exists for fitting an AFT model that incorporates categorical covariates, multiple events, and censored observations simultaneously. This work is dedicated to overcoming the existing challenges by the applications of R programming and data illustration. We arrived at a conclusion that the developed methods are suitable to run and easy to implement in R software. The selection of covariates in the AFT model can be evaluated using model selection criteria such as the Deviance Information Criteria (DIC) and Log-pseudo marginal likelihood (LPML). Various extensions of the AFT model, such as AFT-DPM and AFT-LN, have been demonstrated. The final model was selected based on minimum DIC values and larger LPML values.

**Key words:** censoring, illness-death models, accelerated failure time model, Bayesian Survival Analysis, semi-competing risks.

## 1. Background

Survival analysis is one of the important fields of mathematical statistics and expands to deal with time-to-event data when interest is intended on time and before passing the time an event has occurred, then this kind of data arises. Besides, including statistical methods, it is used for analyzing the time until an event of interest has occurred, where the event is death, the occurrence of any reasonable disease, or other experience of interest. However, we cannot expect each participant to experience the event of interest (like death, cancer) within the study period and get the real data. The prolonged duration of the study period is also affected by censored observation, especially interval censoring, and right censoring.

---

<sup>1</sup>Leicester Real World Evidence Unit, University of Leicester, United Kingdom.  
E-mail: [ab1183@leicester.ac.uk](mailto:ab1183@leicester.ac.uk). ORCID: <https://orcid.org/0000-0002-5757-5513>.

<sup>2</sup>Section of Biostatistics, Centre for Cancer Epidemiology, Tata Memorial Centre, Navi Mumbai 410210, India.  
E-mail: [rajashreedey@gmail.com](mailto:rajashreedey@gmail.com). ORCID: <https://orcid.org/0000-0002-0445-6780>.

We expect subjects to experience only one type of event over follow-up like death from cancer. But in real life, there are so many types of possibilities that subjects can experience more than one type of event in the study period. If death is our interest, then from our observation, we can see that some patients can die from cancer or any traffic accident or in a sudden heart attack. When this kind of event occurs, we refer to these events as "competing events" and the probability of these events as "competing risks." To better understand competing risk scenarios, we can think of a patient who may die from cancer or a heart attack, but he cannot die from both. Sometimes the non-terminal event (like cancer relapse, or readmission) is our research interest. Still, the terminal event (e.g. death) averts the case of the non-terminal event, and it is remarked as semi-competing risk data (Haneuse et al. 2016). Innately we can think of participants of these settings as transitioning through a series of states. For example, we can take cancer relapsing as the non-terminal event and death as the terminal event. Semi-competing risks are inclusive in studies of aging. Here we will give an example for a better understanding of a semi-competing event scenario: a patient who may experience cancer relapse. After some time, he dies of cancer. We can represent the semi-competing risks data in one or more of three transitions: 1) Transition 1: initial condition to the non-terminal event. 2) Transition 2: initial condition to a terminal event. 3) Transition 3: non-terminal events to the terminal event. Semi-competing risks visit the setting where our interest lies to infer a non-terminal event (e.g., disease recurrence, cancer relapse) and a terminal event (e.g., death) and, if possible, for both cases. Let  $T_{i1}$  and  $T_{i2}$  denote time to the non-terminal event and also the terminal event for the  $i^{th}$  study participant respectively. A sturdy association exists between the event's time, so we cannot apply the univariate survival model because it will take the terminal event as an independent event and supply us with overestimated biased results. The semi-competing risks analysis framework appropriately treats the terminal event as a competing event. It considers the dependence between non-terminal and terminal events as a component of the model specification.

The Cox proportional hazard model (Prentice et al. 1992) is used to relate the survival time of a subject to the covariate. We want to find out for which covariate the survival time gets affected. Besides the Cox model, accelerated failure time (AFT) is the essential regression model for censored data (Buckley et al. 1979). The AFT model helps us consider the effect of covariates on survival time. It can offer new insight into risk factors associated with the non-terminal event (cancer relapse) when we conduct such a study among older people, and age is a very relevant factor. In this type of situation, data may have been left truncation. If it is not handled appropriately, each of these situations can give us a biased result of our analysis (Odell et al. 1992). While a statistician or a researcher has so many options for handling these types of situations, there are some works that we have considered. Most of these are on the Cox model for hazard function. About AFT models for semi-competing risk data, there are some recent works, (Dam Ding et al. 2009; Ghosh et al. 2012; Ghosh et al. 2006; Armero et al. 2016; Jiang et al. 2017), but each of them has some limits as they do not consider left-truncation or interval-censoring.

So, this work is dedicated to overcoming the challenges in semi-competing risk data when left-truncation and interval censoring are present. So, we adopt the flexible, study Bayesian framework (Lee et al. 2017) for our analysis of the simulated data as both censorings are adopted in their model. One of the advantages of this framework is we can

take parametric and semi-parametric forms for our baseline survival distribution. We obtained that the developed methods using the functions named **BayesID\_AFT** and **initiate.startValues\_AFT**, which are suitable to run with the help of **SemiCompRisks** (Lee et al. 2015) packages in R.

## 2. Model Framework: Illness death Model

Semi-competing risk data are presented by the participants ready to encounter the two kinds of events and possibly both. We modelled the association between covariates and the two types of event time within the AFT model specification.  $T_{i1}$  and  $T_{i2}$  represent the time of the non-terminal event and also the terminal event for the  $i^{th}$  study participant. Here we adopt the following AFT model specifications (model the times of the events directly) under the illness death modelling framework:

$$\log(T_{i1}) = \mathbf{x}_{i1}^T \boldsymbol{\beta}_1 + \kappa_i + \varepsilon_{i1}, T_{i1} > 0 \tag{1}$$

$$\log(T_{i2}) = \mathbf{x}_{i2}^T \boldsymbol{\beta}_2 + \kappa_i + \varepsilon_{i2}, T_{i2} > 0 \tag{2}$$

$$\log(T_{i2} - T_{i1}) = \mathbf{x}_{i3}^T \boldsymbol{\beta}_3 + \kappa_i + \varepsilon_{i3}, T_{i2} > T_{i1} \tag{3}$$

where  $\mathbf{x}_{ig}$  denotes the vector of transition-specific covariates  $,i = 1, \dots, n$  and  $g \in \{1, 2, 3\}$ .  $\boldsymbol{\beta}_g$  represents the vector of transition specific regression parameters, and  $\varepsilon_{ig}$  denotes the transition-specific random variable whose distribution determines that of corresponding transition time,  $g \in \{1, 2, 3\}$ . Finally,  $\kappa_i$  denotes the random effect of a specific subject in each of (1)-(3) equations that instigates a positive sign of dependency between the two event times.

Let us briefly consider the interpretation of the regression parameter in an AFT model.

From our model given by equation (1), we can write the survivor distribution for the  $i^{th}$  individuals:

$$S_1(t; x_{i1}) = S_{01} \{t \times \exp(-\mathbf{x}_{i1}^T \boldsymbol{\beta}_1)\} \tag{4}$$

where  $S_{01}$  represents the baseline survivor function between the individuals with  $\mathbf{x}_{i1} = 0$ . We can interpret this as a special case when the covariate is dichotomous. Let  $x_{i1}$  be simply a dichotomous covariate with  $x_{i1} = 1$  denoting individuals who have received treatment and  $x_{i1} = 0$  indicates those who did not take treatment. Equation (4) implies that the median time ratio to reason 1 among treatment receiver and non-receiver is  $\exp(\beta_1)$ . For example, if  $x_{i1}$  was a treatment indicator and  $\beta_1 = 0.4$ , we could say that individuals who received the treatment survived 50% longer (as  $\exp(0.4) \approx 1.50$ ) than those individuals who did not receive the treatment, and if  $\beta_1 = -0.4$  then we get  $\exp(-0.4) \approx 0.67$  which indicates 33% shorter survival. If values of  $\exp\{\beta_g\}$  are less than 1.0 for the AFT model, then it indicates an increased risk due to the specific factor. From these results, we can say that a negative value of  $\beta$  suggests increased risk because the event occurs sooner in time.

### 3. Semi-parametric approach in AFT Illness-Death model

There are several works carried on the AFT model through frequency and Bayesian extension (e.g. Christensen et al. 1988; Kuo, L et al. 1997). In this work, we preferred to work with Dirichlet Process Mixture (DPM) prior along with normal distribution for each  $\varepsilon_{ig}$  (Ferguson et al 1973). It helped us to draw  $\varepsilon_{ig}$  independently from a mixture of  $M_g$  with mean and variances from the normal distribution as  $(\mu_{gr}, \sigma_{gr}^2)$ , for  $r \in 1, \dots, M_g$ . Perhaps, it is difficult to identify the distributional form of the  $(\mu_{gr}, \sigma_{gr}^2)$ , so we take each component from normal distribution as being specific to some class and since which class it belongs to is not known to us we prefer to draw from  $G_{g0}$  as a choice of centring distribution. If the 'true' class of membership is not known to us, then  $p_{gr}$  defines the probability to belongs as  $r^{th}$  class for transition  $g$  and  $\mathbf{p}_g = (p_{g1}, \dots, p_{gM_g})$  by the probabilistic representation. It is safe to consider the conjugate symmetric. Dirichlet  $(\tau_g/M_G, \dots, \tau_g/M_G)$  as a choice of prior distribution while the class of memberships for the  $n$  individuals belongs to the  $M_g$  classes, and  $\tau_g$  is presented for the precision parameter. Therefore the mixture distribution is presented as

$$\begin{aligned} \varepsilon_{ig}|r_i &\sim Normal(\mu_{r_i}, \sigma_{r_i}^2), \\ (\mu_{gr}, \sigma_{gr}^2) &\sim G_{g0}, \text{ for } r = 1, \dots, M_g, \\ r_i|\mathbf{p}_g &\sim Discrete(r_i|p_{g1}, \dots, p_{gM_g}), \\ \mathbf{p}_g &\sim Dirichlet(\tau_g/M_G, \dots, \tau_g/M_G). \end{aligned} \tag{5}$$

Now  $M_g \rightarrow \infty$  is presented as DPM along with the normal distribution. This work is presented as Gamma( $a_{\tau_g}, b_{\tau_g}$ ), and hyper-prior for  $\tau_g$ . Now we can take the non-informative flat priors through the real line aligned with the regression line. We can consider  $\kappa_i$  draws from the Normal distribution with  $(0, \theta)$  and finally presented as  $\kappa = \{\kappa_1, \dots, \kappa_n\}$ . Sometimes, we can consider the prior knowledge on the variance component  $\theta$  and adopt the conjugate of the inverse-Gamma hyperprior as IG  $(a_\theta, b_\theta)$ . It is useful to proceed with  $G_{g0}$  as a normal distribution with mean and variance  $\mu_{g0}, \sigma_{g0}^2$ .

### 4. Parametric approach in AFT Illness-Death Model

Sometimes in a small-sample setting parametric-specific model looks more logical as it is easy to handle. For the parametric AFT model, some distributions including Weibull, log-logistic, and log-normal have been proposed for univariate time-to-event data. We consider the log-normal formulation for Bayesian parametric analysis and  $\varepsilon_{ig}$  are taken from an independent normal distribution with mean  $\mu_g$  and variance  $\sigma_g^2$  for  $g \in \{1, 2, 3\}$ . We consider flat priors for location parameters  $(\mu_1, \mu_2, \mu_3)$  on the real line. Independent inverse gamma distributions are considered for  $(\sigma_g^2)$  and denoted as IG  $(a_{\sigma_g}, b_{\sigma_g})$ . We take the same priors for  $\beta_g, \kappa$ , and  $\theta$ , which we took for the DPM model.

### 5. Model comparison criteria

Most of the time, researchers and analysts balanced the compatibility of the specified model with the limitation of the data. In this regard, it is critical to compare the models concerning goodness-of-fit. We used two criteria for this: the deviance information criterion (DIC; Spiegelhalter et al. 2002) and the log-pseudo marginal likelihood statistic (LPML; Geisser et al. 1979). For DIC, we note that work (Celeux et al. 2006) gives a couple of different DIC measures and discusses them in the context of mixture-based random-effects models. In this context, we take their DIC<sub>3</sub> measure based on their guides for our AFT illness-death model given by (1)–(3) and propose the following measure:

$$\begin{aligned}
 DIC_{ID} = & -4E_{\Theta}[\log L(t_{1i}, t_{2i}, \mathcal{D}_i | \Theta) | t_{1i}, t_{2i}, \mathcal{D}_i] \\
 & + 2 \log \prod_{i=1}^n E_{\Theta}[L(t_{1i}, t_{2i}, \mathcal{D}_i | \Theta) | \mathbf{t}_1, \mathbf{t}_2, \{\mathcal{D}_i\}_{i=1}^n]
 \end{aligned}
 \tag{6}$$

In equation (6) all model parameters are denoted by  $\Theta$ , either  $\{\Theta_{SP}, \kappa\}$  or  $\{\Theta_P, \kappa\}$ . In the equation (6) the first term is associated with a deviance that evaluates a goodness-of-fit and the second term computes the measure of complexity. For the purpose of our analysis, we estimate the DIC<sub>ID</sub> with the help of Monte Carlo approximation:

$$\begin{aligned}
 \hat{D}IC_{ID} = & -\frac{4}{Q} \sum_{q=1}^Q \log \left\{ \prod_{i=1}^n L(t_{1i}^{(q)}, t_{2i}^{(q)}, \mathcal{D}_i | \Theta^{(q)}) \right\} \\
 & + 2 \log \left\{ \prod_{i=1}^n \frac{1}{Q} \sum_{q=1}^Q L(t_{1i}^{(q)}, t_{2i}^{(q)}, \mathcal{D}_i | \Theta^{(q)}) \right\}
 \end{aligned}
 \tag{7}$$

At the  $q^{th}$  MCMC iteration,  $\Theta^{(q)}$  denotes the values of  $\Theta$ ,  $q = 1, 2, 3, \dots, Q$ . A model having a smaller DIC value suggests a better fit to the data.

The LPML (2nd comparison criteria) measure is basically the sum of the logarithms subject-specific conditional predictive ordinates and given as  $\sum_{i=1}^n \log CPO_i$ ,

$$\begin{aligned}
 CPO_i = & L(t_{i1}, t_{i2}, \mathcal{D}_i | \{t_{1k}, t_{2k}, \mathcal{D}_k\}_{k \neq i}) \\
 \approx & \left\{ \frac{1}{Q} \sum_{q=1}^Q L(t_{1i}^{(q)}, t_{2i}^{(q)}, \mathcal{D}_i | \Theta^{(q)})^{-1} \right\}^{-1}
 \end{aligned}
 \tag{8}$$

The approximation part in equation (8) pursues from the Monte Carlo estimator (Chen et al. 2012). Note, a model having larger values of LPML suggests a better fit for the data. In this context, one can use the pseudo-Bayes factor (PBF) for the two models by taking the exponent of difference in their LPML values.

### 6. Data

For illustration purposes, we perform some analyses using our simulated data with the primary goal of comparing the two models (parametric & semi-parametric). We simulate

the data consisting of  $n=5000$  on the frame of semi-competing risk data where the interest lies in a non-terminal event that is subject to a terminal event which is a competing risk for the non-terminal event but not vice versa. We have developed this kind of semi-competing risk data frame with five dichotomous covariates and fit the model. Table 1 represents the baseline characteristic (sex, race, etc.) of 5000 participants. We adopted interval censoring and left truncation also. It also provides a 60-months summary of outcomes, overall and within levels of the factors reported. From the first row of Table 1, we can see that 25.58% participants are censored for both events and 25.24% experienced both events. We also see that a total of 1239 individuals experienced the non-terminal event and were censored for the terminal event, and 1220 participants have experienced the terminal event without having the non-terminal event.

Beyond the overall rates, Table 1 reveals substantial variation in the distribution of the four outcome types across levels of certain factors. We see, for example, that the rates at which individuals have experienced both events within 60 months is 26.02% among the individuals having  $I_{(1)} = 1$  to 24.03% among individuals having  $I_{(1)} = 0$ .

Table 1: Overall information about covariates based on 5000 individuals experienced on non-terminal and/or terminal events.

	Total n(%)	censored n(%)	Non-terminal event only n(%)	Terminal event only n(%)	Both events n(%)
Total	5000 (100)	1279 (25.58)	1239 (24.76)	1220 (24.4)	1262 (25.24)
Covariate 1					
$I_{(1)} = 1$	3032 (60.64)	740 (24.41)	757 (24.97)	746 (24.60)	789 (26.02)
$I_{(1)} = 0$	1968 (39.36)	539 (27.39)	482 (24.49)	474 (24.09)	473 (24.03)
Covariate 2					
$I_{(2)} = 1$	4532 (90.64)	1156 (25.51)	1123 (24.78)	1109 (24.47)	1144 (25.24)
$I_{(2)} = 0$	468 (9.36)	123 (26.28)	116 (24.79)	111 (23.72)	118 (25.21)
Covariate 3					
$I_{(3)} = 1$	3896 (77.92)	991 (25.44)	983(25.23)	968 (24.84)	954 (24.49)
$I_{(3)} = 0$	1104 (22.08)	228 (20.65)	256 (23.19)	252 (22.83)	288 (26.09)
Covariate 4					
$I_{(4)} = 1$	2550 (51)	649 (25.45)	637 (24.98)	648 (25.41)	616 (24.16)
$I_{(4)} = 0$	2450 (49)	630 (25.71)	602 (24.57)	572 (23.35)	646 (26.37)
Covariate 5					
$I_{(5)} = 1$	2789 (55.78)	701 (25.13)	690 (24.74)	692 (24.81)	706 (25.31)
$I_{(5)} = 0$	2211 (44.22)	578 (26.14)	549 (24.83)	528 (23.88)	556 (25.15)

## 7. Results

### 7.1. Overall model fit

Table 2 provides the calculated values obtained on AFT-LN and AFT-DPM by corresponding DIC and LPML values. The models, i.e., AFT-LN and AFT-DPM, considered the  $\kappa_i$  (random effect) and resulted in DIC values as 34810 and 30975. So, AFT-DPM is relevant in this data analysis context compared to the AFT-LN model. Similarly, the second measure (LPML) obtained on AFT-LN and AFT-DPM is -16156 and -15719. It also confirms that AFT-DPM is more relevant than AFT-LN.

Table 2: DIC and LPML for two proposed models fit to simulated data.

	DIC	LPML
AFT-LN	34810	-16156
AFT-DPM	30975	-15719

### 7.2. Analysis: Covariate effect

As mentioned earlier in Section 3, if values of  $\exp\beta_g$  are less than 1.0 for the AFT model, then it indicates an increased risk due to the specific factor. Table 3 presents the posterior median (PM) and 95% credible interval for the regression parameter from our analysis having a patient-specific random effect.

From the first column of Table 3, we have found proof that individuals with 0 indicators for 1st covariate  $I_{(1)}$  significantly increased for reason 1 for the AFT-DPM analyses. The median time to reason 1 is estimated to be 6.5% shorter for these individuals than those with one indicator for  $I_{(1)}$ .

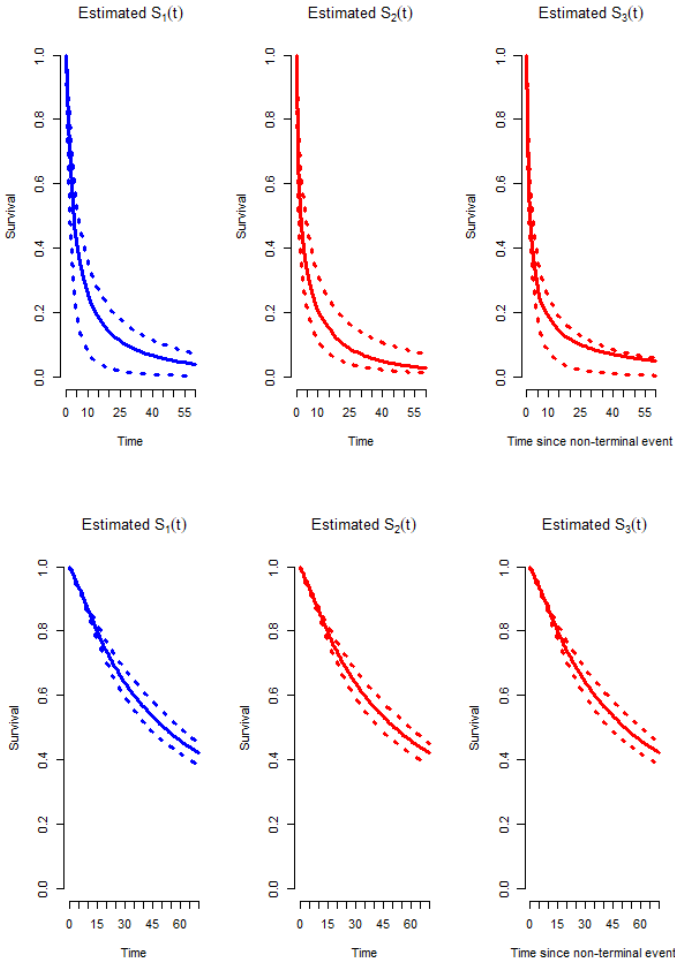
We also find that individuals with 0 indicators for 2nd covariate  $I_{(2)}$  have a lower risk of reason 3. Their median time to reason 3 (non-terminal to terminal event) is estimated to be 64.7% and 7.6% longer than individuals who have the one indicator under the AFT-LN and AFT-DPM model, respectively.

Table 3: Estimated posterior medians along with 95% credible intervals (CI) for transition  $g = 1, 2, 3$  based on two type of AFT-illness death model including the patient-specific random effect.

	Transition 1 (95% CI)	Transition 2(95% CI)	Transition 3 (95% CI)
Covariate 1			
AFT-LN	1.432(1.235,2.280)	2.719(2.072,3.758)	1.578(1.468,2.351)
AFT-DPM	0.935 (0.893,0.966)	1.079(1.008,1.151)	1.082(1.053,1.108)
Covariate 2			
AFT-LN	3.070(2.626,3.429)	2.080(1.828,2.659)	1.647(1.262,2.493)
AFT-DPM	1.127(1.099,1.155)	1.296(1.110,1.369)	1.076(0.984,1.087)
Covariate 3			
AFT-LN	1.991(1.709,2.262)	2.385(2.129,4.117)	2.464(1.928,2.853)
AFT-DPM	1.045(1.031,1.105)	1.009(0.925,1.038)	1.079(1.056,1.149)
Covariate 4			
AFT-LN	1.725(1.402,2.384)	1.758(1.602,2.351)	2.275(1.786,4.046)
AFT-DPM	1.058(1.058,1.058)	0.975(0.890,0.997)	1.208(1.042,1.851)
Covariate 5			
AFT-LN	1.604(1.389,1.721)	2.697(1.596,3.605)	2.028(1.471,2.339)
AFT-DPM	1.073 (1.070,1.075)	1.181(0.895,1.417)	1.191(1.145,1.237)

## 8. Discussion

In this work, we discuss the semi-competing risks framework as a way of investigating variation in risk for a non-terminal event where the occurrence of the event is subject to a



Estimated survival function for the three transitions specific survival distribution based on the AFT-LN & AFT-DPM model respectively.



terminal event. In this context, we have analyzed the semi-competing risk data using the proposed AFT illness death model (Lee et al. 2017), which serves as a helpful complement of the traditional hazard-based model of say (Xu et al. 2010) and (Lee et al. 2015).

Crucially the two modelling frameworks characterize associations through fundamentally different contrasts (see Section 2.2) and, in this sense, jointly provide an expanded scope for scientific inquiry. As such, reckoning on the scientific background and goals, analysts may value more highly to consider one or the other or possibly both.

In this text, our main objective is to find which model is a better fit for our frame and to estimate the effects of the covariates on the risk of the non-terminal event (e.g. cancer relapse). At the same time, we assume that death plays a vital role in the analysis. We have handled the data carefully as left truncation and interval censoring are present. If we do not consider these things, we will get a biased result.

In this article, we build the framework through the Bayesian model, which will help the researchers to take the advantage of well-known benefits including the ability to naturally incorporate prior information and the automated quantification of prediction and uncertainty.

Finally, we note that there are a number of ways in which one could build the proposed framework. First, while the focus of this article has been on semi-competing risk data, we have developed and implemented analogous parametric and semi-parametric univariate AFT models in settings where left truncation and interval censoring are present. Such a model might, for example, be useful if interest lies in whether there are differences in mortality between patients with and without a diagnosis of Alzheimer's and dementia. Some specific areas where the model can be used include pregnancy, where delivery is the terminal event and Preeclampsia will be the non-terminal event and palliative care where death is the terminal event and readmission will be the non-terminal event.

## **Ethical statement**

Authors consciously assure that the following is fulfilled for the manuscript:

- 1) This material is the authors' own original work, which has not been previously published elsewhere.
- 2) The paper is not currently being considered for publication elsewhere.
- 3) The paper reflects the authors' own research and analysis in a truthful and complete manner.
- 4) The paper properly credits the meaningful contributions of co-authors and co-researchers.
- 5) The results are appropriately placed in the context of prior and existing research.
- 6) All sources used are properly disclosed (correct citation). Literally copying of text must be indicated as such by using quotation marks and giving proper reference.
- 7) All authors have been personally and actively involved in substantial work leading to the paper, and will take public responsibility for its content.

## **Conflict of interest**

The authors declared there is no conflict of interest.

## Acknowledgement

The author would like to acknowledge the support and resources provided by the Section of Biostatistics, Centre for Cancer Epidemiology, Tata Memorial Centre's staff, administrators, and teachers who work tirelessly to ensure their students receive the highest quality education.

## References

- Adam Ding, A., Shi, G., Wang, W. and Hsieh, J. J., (2009). Marginal regression analysis for semi-competing risks data under dependent censoring. *Scandinavian Journal of Statistics*, 36(3), pp. 481–500.
- Armero, C., Cabras, S., Castellanos, M. E., Perra, S., Quirós, A., Oruezábal, M.J. and Sánchez-Rubio, J., (2016). Bayesian analysis of a disability model for lung cancer survival. *Statistical methods in medical research*, 25(1), pp. 336–351.
- Buckley, J., James, I., (1979). Linear regression with censored data. *Biometrika*, 66(3), pp.429-436.
- Celeux, G., Forbes, F., Robert, C. P. and Titterton, D. M., (2006). Deviance information criteria for missing data models. *Bayesian analysis*, 1.4 (2006), pp. 651–673.
- Chen, M. H., Shao, Q. M. and Ibrahim, J. G., (2012). Monte Carlo methods in Bayesian computation. *Springer Science & Business Media*.
- Christensen, R., Johnson, W., (1988). Modelling accelerated failure time with a Dirichlet process. *Biometrika*, 75(4), pp. 693–704.
- Ferguson, T. S., (1973). A Bayesian analysis of some nonparametric problems. *The annals of statistics*, pp. 209–230.
- Geisser, S., Eddy, W. F., (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365), pp. 153–160.
- Ghosh, D., Taylor, J. M. and Sargent, D. J., (2012). Meta-analysis for surrogacy: Accelerated failure time models and semicompeting risks modeling. *Biometrics*, 68(1), pp. 226–232.
- Ghosh, S. K., Ghosal, S., (2006). Semiparametric accelerated failure time models for censored data. *Bayesian statistics and its applications*, 15, pp. 213–229.

- Haneuse, S., Lee, K.H., (2016). Semi-competing risks data analysis: accounting for death as a competing risk when the outcome of interest is nonterminal. *Circulation: Cardiovascular Quality and Outcomes*, 9(3), pp. 322–331.
- Jiang, F., Haneuse, S., (2017). A semi-parametric transformation frailty model for semi-competing risks survival data. *Scandinavian Journal of Statistics*, 44(1), pp.112–129.
- Kuo, L., Mallick, B., (1997). Bayesian semiparametric inference for the accelerated failure-time model. *Canadian Journal of Statistics*, 25(4), pp.457–472.
- Lee, K.H., Haneuse, S., Schrag, D. and Dominici, F., (2015). Bayesian semi-parametric analysis of semi-competing risks data: investigating hospital readmission after a pancreatic cancer diagnosis. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, 64(2), p. 253.
- Lee, K. H., Lee, C., Alvares, D., Haneuse, S. and Lee, M. K. H., (2015). Package ‘Semi-CompRisks’.
- Lee, K. H., Rondeau, V. and Haneuse, S., (2017). Accelerated failure time models for semi-competing risks data in the presence of complex censoring. *Biometrics*, 73(4), pp. 1401–1412.
- Odell, P. M., Anderson, K. M. and D’Agostino, R. B., (1992). Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics*, pp. 951–959.
- Prentice, R. L., (1992). Introduction to Cox, 1972, regression models and life-tables. *Breakthroughs in Statistics: Methodology and Distribution*, pp. 519–526.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A., (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4), pp. 583–639.
- Xu, J., Kalbfleisch, J. D. and Tai, B., (2010). Statistical analysis of illness–death processes and semicompeting risks data. *Biometrics*, 66(3), pp. 716–725.



# Census administration in Ukraine: insight into the Polish experience in the context of international indicators analysis

Svitlana Chugaievska<sup>1</sup>, Grażyna Dehnel<sup>2</sup>, Andrey Targonskii<sup>3</sup>

## Abstract

The latest National Population Census in Poland, like in many EU countries, was conducted in 2021. In Ukraine, during the entire period of independence, a national census was conducted only once, in 2001, while the following rounds kept being postponed. In 2019, a trial census was conducted in several regions of central Ukraine. The working hypothesis is that for the organisation and conduct of the All-Ukrainian Population Census in Ukraine, it is important to use the experience of EU countries in this field (and Poland's experience in particular). The purpose of the article is to substantiate the theoretical foundations and comparative analysis of the processes of conducting censuses in Poland and Ukraine and to study the factors of increasing the level of respondent participation. The article reviews a number of socio-economic factors in the light of the comparison of data census organisation in Poland and Ukraine. Several international indicators were investigated that could have an impact on respondent participation in statistical surveys and censuses. By applying factor analysis, it was possible to identify the factors that could increase the level of respondent participation. To solve these demographic challenges, the following activity should be considered: financial support for the successful functioning of economic entities, improvement of social infrastructure, friendly environment for innovation and investment, and peace and political stability in Ukraine.

**Key words:** national census, trial census, international comparisons, multifactorial statistical analysis of the civic activity level.

## 1. Introduction

The role and significance of censuses remain the focus of attention for many researchers, economists and politicians. According to UN standards, a national census

---

<sup>1</sup> Department of Economics and Innovation, Jagiellonian University in Krakow, Poland & Zhytomyr Ivan Franko State University, Department of Mathematical Analysis, Business Analysis and Statistics, Ukraine. E-mail: [schugaevskaya@ukr.net](mailto:schugaevskaya@ukr.net). ORCID <https://orcid.org/0000-0002-9751-9647>.

<sup>2</sup> Department of Statistics, Poznan University of Economics and Business, Poland. E-mail: [grazyna.dehnel@ue.poznan.pl](mailto:grazyna.dehnel@ue.poznan.pl). ORCID <https://orcid.org/0000-0002-0072-9681>.

<sup>3</sup> Zhytomyr Ivan Franko State University, Department of Mathematical Analysis, Business Analysis and Statistics, Ukraine. E-mail: [targonsk@zu.edu.ua](mailto:targonsk@zu.edu.ua). ORCID <https://orcid.org/0000-0002-0670-037X>.



should be conducted in each country every 10 years. This recommendation is observed in EU countries. The most recent population census in Poland, like in many other EU members, was conducted in 2021, whereas in Ukraine – in 2001. Over the past 21 years, there have been a number of attempts to hold another census in Ukraine, but each time the date was postponed.

However, at the end of 2019, a trial census was conducted in several districts of the Kyiv region. The trial census relied on the latest digital technology, enabling respondents to fill out the census questionnaire via a smartphone app. Direct interviews were also conducted. But the main problem of the trial census was that the realized sample accounted for just 75% of the planned sample, with less than 1% collected by means of the app.

A timely and reliable administration of a national census and its statistical outputs are crucial for the implementation of the main directions of the country's sustainable development, especially the prospects of EU membership. This development is associated with three main goals: overcoming poverty, improving food security indicators, improving the well-being of the population. It should be emphasized that all these goals require information about the number of permanent residents in particular regions and in the country as a whole. In addition, once the hostilities in Ukraine are over, the most urgent task of the authorities will be to restore territorial integrity and start the reconstruction of Ukrainian cities and villages destroyed by the war. Such investment plans will require detailed information about the population in each region.

On June 23, 2022, Ukraine acquired the status of a candidate for EU membership. As a result, the government is now faced with the difficult task of adapting the national laws and standards to the European legislation, ensuring compliance with the rule of law, embracing democratic values and promoting the market economy. In order to continue reforms with a view to becoming an EU member, Ukraine needs to conduct regular population censuses according to EU standards. In addition, reliable and standardized national statistics are necessary for international comparisons made by the UN, the World Bank, the International Bank for Reconstruction and Development, Eurostat, the International Monetary Fund and other world organizations.

Since the declaration of independence in 1991, official statistics in Ukraine has undergone significant changes and the independence and objectivity of statistical information has become the main objective underlying the reforms undertaken in the organization. Today, the task of official statistics is not only to provide objective socio-economic indicators at the macro and micro levels, but also to contribute to the democratization of society by ensuring the sustainable development of all regions (Osaulenko O., et al., 2021).

In recent years, a number of researchers have focused on the trends and challenges connected with the demographic situation in the countries of Central and Eastern

Europe (Krywult-Albańska M., 2012, Marciniak G., 2014, Da Costa J. N., Bielecka E., Calka B., 2017). Dygaszewicz J. (2020) calls for a more extensive use of modern IT technologies in the organization of population censuses and data from administrative registers. Some authors point to differences between countries regarding the conditions that are relevant for conducting national censuses, which are related to significant differences between their national economies. Wisła R. et al. (2020) compare structural changes in the Polish and Ukrainian economies with those observed in other Central and Eastern European countries.

Zayukov I. (2011), Kravchenko V. and Kravchenko N. (2015) emphasize the need to study the causes and consequences of the deep demographic crisis in Ukraine. Libanova E., 2013, Melnik S., 2014 and Malish N., 2016 highlight the need to solve demographic problems in order to ensure social development of different regions. However, the problems associated with the organization of a national population census and other statistical surveys, especially now in the conditions of martial law, remain unresolved.

Since the latest census data are over 20 years old, information about the population in the country's regions comes only from the Civil Registration of the Population Office and the State Migration Service. Because of high rates of migration, including refugee and labour migration, the accurate measurement of the population remains a challenge (Libanova E., 2019, Malynovska O., 2016). Since 2014, the country has experienced high levels of refugee migration resulting from political processes and Russia's invasion of the Crimean Peninsula. Migration flows intensified following Russia's aggression against Ukraine on February 24, 2022. As war hostilities escalated, large groups of the population were forced to leave their homes and move to neighbouring countries: Poland, Slovakia, Germany, Bulgaria, the Czech Republic, etc. (Kolomiets O., 2022).

Will these people be able to return home and if so, when? Will they be able to participate in the census? How will it be possible to organize and conduct a census given such a high level of migration and political instability in the country? What experiences of the EU countries regarding census administration can be used to help Ukraine overcome these challenges?

The authors believe that in order to conduct a successful population census in Ukraine, it is important to implement the experience of the EU countries, particularly Poland, regarding the organization of censuses and household surveys. The main results of the study were presented at the 40th International Conference MSA-2022, Lodz, November 7-10, 2022<sup>4</sup>. The authors are sincerely grateful to the organizers of the Conference for the high evaluation of the research and the recommendations.

The purpose of the article is to provide the theoretical foundations and compare the processes of conducting censuses in Poland and Ukraine, with emphasis on factors

---

<sup>4</sup> <https://sites.google.com/view/msa-lodz>

that increase respondent participation. The authors compare the 2021 census in Poland and the 2019 trial census in Ukraine, by analyzing census forms, effectiveness indicators of various response modes: online, mail, telephone or personal interview, indices of the socio-economic development of the countries, which can stimulate respondent participation.

## 2. Data and methodology

The following comparative analysis is based on publicly available data published online by national statistical institutes in Poland<sup>5</sup> and Ukraine<sup>6</sup>.

A comparative analysis of the census forms composition and respondent participation rates is based on information included in methodological reports accompanying each census. Factors that can affect the degree of respondent participation were selected from among international economic indices for each country, published by the World Bank<sup>7</sup>, the European Commission and Eurostat<sup>8</sup>, the International Labor Organization<sup>9</sup> and the World Economic Forum<sup>10</sup>.

When selecting data to describe respondent participation in both countries, three groups of indices were considered:

- 1) indicators relating to electronic document flow: E-Participation Framework Index (EPFI)<sup>11</sup> and the UN Global E-Government Development Index (EGDI)<sup>12</sup>. The first characterizes the use of state electronic services, the second – the use of electronic services in state document circulation.
- 2) indicators of the country's sustainable economic development: Fragile States Index (FSI)<sup>13</sup>, Global Innovation Index (GII)<sup>14</sup> and The Economic Complexity Index (ECI)<sup>15</sup>. The first one characterizes the instability of state institutions, weak protection of the population, lack of access to medical and educational services, etc. The second measures the level of innovative processes, such as the use of new technologies, energy production and sustainable products, etc. The third index is based on the diversity and complexity of their export basket. It reveals the diversity

---

<sup>5</sup> <https://stat.gov.pl/en>

<sup>6</sup> <http://www.ukrstat.gov.ua>

<sup>7</sup> <https://data.worldbank.org/>

<sup>8</sup> <https://ec.europa.eu/eurostat/data/database>

<sup>9</sup> <https://www.ilo.org/global/about-the-ilo/newsroom/lang--en/index.htm>

<sup>10</sup> <https://www.weforum.org/reports/>

<sup>11</sup> <https://publicadministration.un.org/egovkb/en-us/About/Overview/E-Participation-Index>

<sup>12</sup> <https://publicadministration.un.org/egovkb/en-us/About/Overview/-E-Government-Development-Index>

<sup>13</sup> <https://fragilestatesindex.org/indicators/>

<sup>14</sup> <https://www.globalinnovationindex.org/Home>

<sup>15</sup> <https://oec.world/en/rankings/eci/hs6/hs96?tab=table>



and sophistication of the productive capabilities embedded in the exports of each country.

- 3) social indicators: Human Development Index (HDI)<sup>16</sup> and Social Progress Index (SPI)<sup>17</sup>. The first is a measure of average achievements in key dimensions of human development: a long and healthy life, education and decent standards of living. The second one is an aggregate country score with respect to three dimensions: basic human needs, foundations of well-being and opportunities. It relies exclusively on social and environmental indicators and measures outputs not inputs.

Guided by the principle of data comparability, the authors chose a single common series of data regarding the selected international indices for Poland and Ukraine – from 2009 to 2021. However, it turned out that some of these indices are calculated once every two years, which means no data were available for some years. For some indices, information for 2021 has not been published yet. In such cases, the authors applied a method of working with missing data developed by Fichman M. (2003). The empty cells were filled with values consistent with the general trend observed for a given index and the application of regression equations.

Sociometric, economic and statistical methods were used in the analysis. By comparing absolute, relative and average indicators of realized sample size with the planned number of respondents, it was possible to determine response rates for personal interviews and the electronic mode of response (Mokin B., Mokin O., 2015).

In the assessment of census forms used in Poland and Ukraine, the authors selected 8 groups of questions: 1) demographic characteristics (age, gender, place of birth, etc.); 2) ethnic origin, language, religion, citizenship; 3) education; 4) employment; 5) migration activity; 6) living conditions; 7) family connections; 8) health characteristics, disability status.

The method of principal components analysis (PCA) was used to identify the most important factor affecting respondent participation in national surveys (Dunteman G., 1989). The PCA method made it possible to reduce the size of the initial database in order to select factors with the greatest importance for the issue of interest – the level of respondent participation. The method was applied in the following stages.

*In the first step*, data approximation was performed by linear images, where the Euclidean distance between each of the vectors and its linear image is minimized. In this study, the source dataset includes a finite set of vectors:  $I_1, I_2, \dots, I_m \in R^n$  for each  $k = 1, 2, \dots, n - 1$  among all  $k$ -dimensional linear images one must find one where

---

<sup>16</sup> <https://ourworldindata.org/human-development-index>

<sup>17</sup> <https://www.socialprogress.org>

$L_k \in R^n$  and the sum of the squares of the Euclidean distances from each vector to the linear image is minimal:

$$\sum_{i=1}^7 \text{dist}^2(x_i, L_k) \rightarrow \min \quad (1)$$

On the other hand, when any  $k$ -dimensional linear image in the  $R^n$  space can be represented as a set of linear combinations:

$$L_k = \{a_0 + a_1\beta_1 + a_2\beta_2 + \dots + a_k\beta_k\}, \quad (2)$$

where  $\beta_i$  denotes some parameters of this linear combination,  $a_0$  is a free element and  $\{a_1; a_2; \dots; a_k\}$ , which belong to the  $R^n$  space, is called an orthonormal set of vectors or vectors of principal components.

In this case, we present the sum of the squares of the Euclidean distances as the Euclidean norm:

$$\sum_{i=1}^7 \text{dist}^2(x_i, L_k) = \|x_i - a_0 - \sum_{j=1}^k a_j ((a_j, x_i) - a_0)\|^2 \quad (3)$$

The solution of this approximation problem was to find a series of nested linear images  $L_0 \subset L_1 \subset L_2 \dots \subset L_{n-1}$ , where:

$$L_k = \{a_0 + a_1\beta_1 + a_2\beta_2 + \dots + a_k\beta_k\},$$

and these images were determined by a set of vectors' main components.

*In the second step*, we looked for orthogonal projections with the largest value of dispersion. The first principal component was selected, where the sample variance of the data on the first coordinate is maximal. Next, the second main component was selected, where the sample variance along the second coordinate is maximal, provided that it is orthogonal to the first coordinate. And so on for the  $k$ -th principal component. Each sample variance for the  $k$ -th principal component in this study was determined by the formula:

$$S^2_{\max}(a_k, x_i) = \frac{1}{m} \sum_{k=1}^m (a_k, x_i)^2 \quad (4)$$

*In the third step*, we looked for orthogonal projections with the largest root mean square distance between points. This enabled us to compare and weight different pairwise distances between indices of respondent participation.

*In the last, fourth step*, correlations between the coordinates and indices used were cancelled. That is, we selected only those main components for which the coefficient of covariance between their various coordinates was equal to zero. This enabled us to select the factors which are relevant to conduct of a census and increase the level of respondent participation.

The method of scientific generalization was used to develop directions for improving the administration of a population census in Ukraine as soon as military operations in the country end. Results of the factor analysis of international economic indices were used to propose directions for improving respondent participation in statistical surveys.

### **3. Socio-economic conditions of relevance for census administration in Poland and Ukraine**

Before one draw on the Polish experiences of census administration to inform Ukrainian reforms in this respect, it is necessary to note a significant difference between the two countries. Poland's accession to the EU in 2004 was an important moment in the process of transformation. Today, Poland is one of the most dynamically developing economies in the EU and ranks quite high in terms of the Human Development Index. It is a country with relatively high indicators of income and the quality of life, the level of security, the quality of education and economic freedom. In 2000, Poland's nominal GDP per capita was USD 4,501, whereas in 2021 it was USD 17,840, which represents a 4-fold increase. As a result, Poland was 44th in the ranking of countries by GDP per capita. In the same period, Ukraine's GDP per capita rose from USD 636 to 4835 USD, i.e. 7.6 times. However, in terms of GDP per capita, Ukraine ranks 101st in the world and is the "poorest country in Europe". According to the preliminary results of the 2021 census, the population of Poland is over 38.5 million, making it the fifth most populous country in the EU, and the eighth in Europe. It is also the ninth biggest country in Europe.

Currently, Ukraine is an industrial-agrarian economy, with a predominance of raw material production. The country is one of the leading exporters of many types of agricultural products. Major sectors of the Ukrainian economy include the mining industry, separate branches of mechanical engineering, ferrous and non-ferrous metallurgy, etc. Ukraine is an important producer of electricity as well as military equipment and weapons. Although the country ranks 74th in terms of the Human Development Index, (for comparison, Poland is 33rd), the standard of living and indicators of the quality of life vary greatly for different categories of the population. Other problems the country faces include a high level of corruption as well as poor security and legal protection of citizens. Since February 2014, Ukraine has been defending itself against the armed invasion of the Russian Federation, which led to the annexation of Crimea and the occupation of parts of the Donetsk and Luhansk regions. The next stage of the Russian-Ukrainian war began on February 24, 2022, with Russia's large-scale invasion of Ukraine. The war has contributed to a strong growth of patriotic sentiments among Ukrainians, which is reinforced by humanitarian aid and military support provided by the international community. According to the 2001 census, the Ukrainian population was 48.5 million, and according to estimates based on the population register, it was 41.3 million at the start of 2021.

The authors believe that the low level of non-response bias is correlated with high standards of living when part of the information comes from administrative databases.

Therefore, having considered Poland's experience in organizing and conducting censuses, the authors singled out 3 groups of international indices:

- 1) indicators characterizing the degree of electronic document circulation;
- 2) indicators characterizing the stability of the country's economic development;
- 3) indicators characterizing the standard of living (Table 1).

Considering these indices, the situation of Poland looks much more favourable compared to that of Ukraine. In the case of the FSI index, the higher the country's position in the ranking, the more fragile it is.

**Table 1.** Socio-economic factors of relevance for respondent participation a year before the national census in Poland and the trial census in Ukraine

Indices	Country's rank for a given index	
	Poland, 2020	Ukraine, 2018
<i>Indicators of electronic document flow</i>		
1. E-Participation Framework Index (EPFI) <sup>18</sup> , I1	9	75
2. The UN Global E-Government Development Index (EGDI) <sup>19</sup> , I2	24	82
<i>Indicators of sustainable economic development</i>		
3. Fragile States Index (FSI) <sup>20</sup> , I3	145	86
4. Global Innovation Index (GII) <sup>21</sup> , I4	38	43
5. The Economic Complexity Index (ECI) <sup>22</sup> , I5	24	41
<i>Social indicators</i>		
6. Human Development Index (HDI) <sup>23</sup> , I6	35	78
7. Social Progress Index (SPI) <sup>24</sup> , I7	31	64

Source: based on data published on the official websites listed in the footnotes.

The E-Participation Framework Index (EPFI), included in the first group of indicators, is a complementary index to the UN E-Government Survey. It focuses on the use of online services that governments use to facilitate the provision of information to citizens (e-information exchange), engagement with stakeholders (e-consultation)

<sup>18</sup> <https://publicadministration.un.org/egovkb/en-us/About/Overview/E-Participation-Index>

<sup>19</sup> <https://publicadministration.un.org/egovkb/en-us/About/Overview/-E-Government-Development-Index>

<sup>20</sup> <https://fragilestatesindex.org/>

<sup>21</sup> <https://www.globalinnovationindex.org/Home>

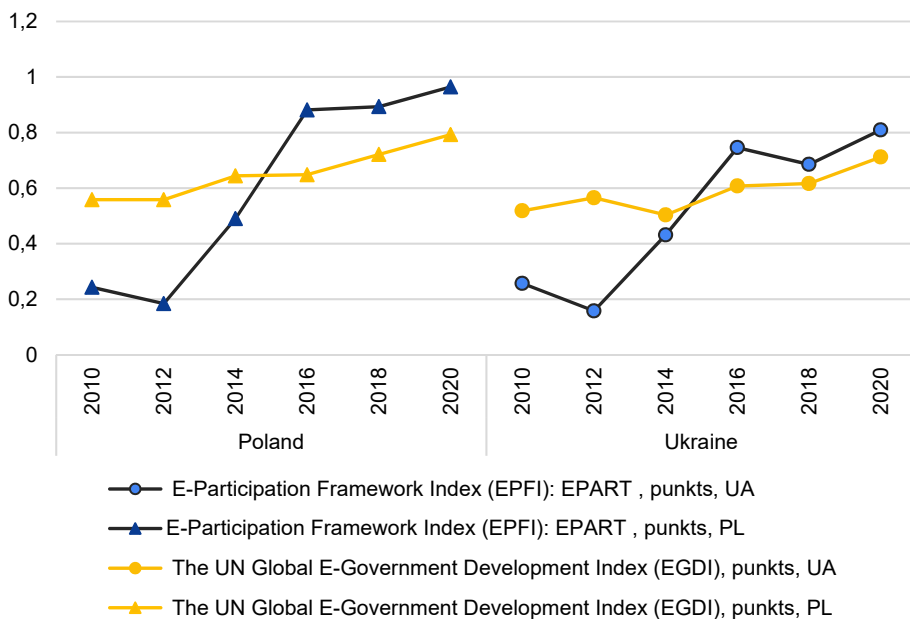
<sup>22</sup> <https://oec.world/en/rankings/eci/hs6/hs96?tab=table>

<sup>23</sup> <https://ourworldindata.org/human-development-index>

<sup>24</sup> <https://www.socialprogress.org>

and participation in decision-making processes (e-decision-making). It includes the following sub-indexes: framework system of e-participation (the degree of citizens' e-participation in various events organized by the state), electronic information ("enabling participation by providing citizens with public information and access to information without or upon demand"), electronic consultation ("engaging citizens in contributions to and deliberation on public policies and services") and electronic decision-making ("empowering citizens through co-design of policy option and co-production of service components and delivery modalities").

It should be noted that in 2010, the values of the EPI were almost the same for both countries (0.249 for Poland and 0.2571 for Ukraine). However, in the following years, e-participation of the Polish population improved, especially between 2014 and 2016, when the value of the index almost doubled from 0.4902 to 0.8814, which is reflected by a jump in the world ranking from 65th to 14th place. In 2020, the EPI for Poland was 0.9643, while the corresponding value for Ukraine was 0.8095 (Figure 1).



**Figure 1.** A comparison of indicators of electronic document flow in Poland and Ukraine, 2010–2020  
 Source: based on data published on the official websites listed in the footnotes<sup>25</sup>.

<sup>25</sup> <https://publicadministration.un.org/egovkb/en-us/About/Overview/E-Participation-Index>;  
<https://publicadministration.un.org/egovkb/en-us/About/Overview/-E-Government-Development-Index>

The second indicator of electronic document flow is the UN Global E-Government Development Index (EGDI). In addition to assessing the website development patterns in a country, the EGDI considers access characteristics, such as the infrastructure and educational levels, to reflect how a country is using information technologies to promote access and inclusion of its people. The EGDI is a composite measure of three dimensions of e-government: provision of online services (Online Services Index, OSI), development status of telecommunication infrastructure (Telecommunication Infrastructure Index, TII), and inherent human capital (Human Capital Index, HCI).

The values of the EGDI confirm the pattern observed in the case of the EPI. While in 2010 its respective values for Poland and Ukraine were approximately the same (0.5582 vs 0.5181), a decade later the situation in Poland improved considerably: the index risen to 0.7986, but at the same time the country dropped from the 45th place in 2010 to the 24th in 2020. As for Ukraine, because of political instability and Russia's military aggression, the evident improvement from 0.5181 in 2010 to 0.7119 in 2019 was not big enough to secure a better position in the ranking – the country actually dropped from 54th to 69th place.

In addition to having an efficient system of e-participation and e-government, which can have an impact on respondent participation in statistical surveys, another group of indicators is connected with the development of a competitive national economy and the quality of life.

The first indicator in this group is the Fragile States Index, which is based on a Conflict Assessment System Tool (CAST) developed by the Fund for Peace (FFP) nearly a quarter of a century ago to assess “vulnerabilities which contribute to the risk of state fragility”<sup>26</sup>. The CAST framework was originally developed to measure these vulnerabilities and assess how they might affect projects in the industry, and continues to be widely used by policy makers, field practitioners and local community networks. An unstable situation in a state can have serious consequences not only for that state and its population, but also for its neighbours and other countries elsewhere in the world. Internal conflicts, humanitarian and political crises can arise from ethnic tensions; some turn into civil wars; others take the form of revolutions, and lead to complex humanitarian emergencies.

The Fragile States Index (FSI) is “is a critical tool in highlighting not only the normal pressures that all states experience, but also in identifying when those pressures are outweighing a states' capacity to manage those pressures”. The FSI is based on 12 indicators of the CAST framework, which are grouped into 4 categories<sup>27</sup>: Cohesion Indicators (C1 – Security apparatus, C2 – Fractional elites, C3 – Group grievance); Economic Indicators (E1 – Economic decline, E2 – Uneven economic development,

---

<sup>26</sup> <https://fragilestatesindex.org/methodology/>

<sup>27</sup> <https://fragilestatesindex.org/indicators/>

E3 – Flight of people and brain drain); Political Indicators: (P1 – Legitimacy of the state, P2 – Public services, P3 – Human rights and the rule of law) and Social and Interdisciplinary Indicators (S1 – Demographic pressure, S2 – Refugees and IDPs, X1 – External intervention). Unlike the previous indices, a higher value indicates a more fragile state, with a more unstable economy.

In terms of the FSI, Ukraine's situation deteriorated from 69.7 in 2009 to 91.0 in 2021, which is reflected by the higher position in the ranking of fragile states: from 110<sup>th</sup> to 69<sup>th</sup> place, indicating high instability and the presence of conflicts in society. This is the result of two political revolutions in 2008 and 2014, as well as Russia's military aggression followed by the annexation of parts of the Ukrainian territory. The situation of Poland has been considerably more stable, as evidenced by the decline in the ranking from 49.6 (142<sup>nd</sup> place) in 2009 to 43.1 in 2021 (147<sup>th</sup> place) (Figure 2).

The second indicator of economic development is the Global Innovation Index (GII), which tracks the latest global innovation trends. Every year, it evaluates the efficiency and innovativeness of world economies. It also highlights indicator strengths and weaknesses of each country to give a more detailed description of its innovation activities. Currently, the overall GII ranking is based on 81 indicators grouped into two sub-indices: Innovation Input Sub-Index, consisting of 5 pillars (including measures of political environment, education, infrastructure and knowledge creation) and Innovation Output Sub-Index, consisting of two pillars<sup>28</sup>.

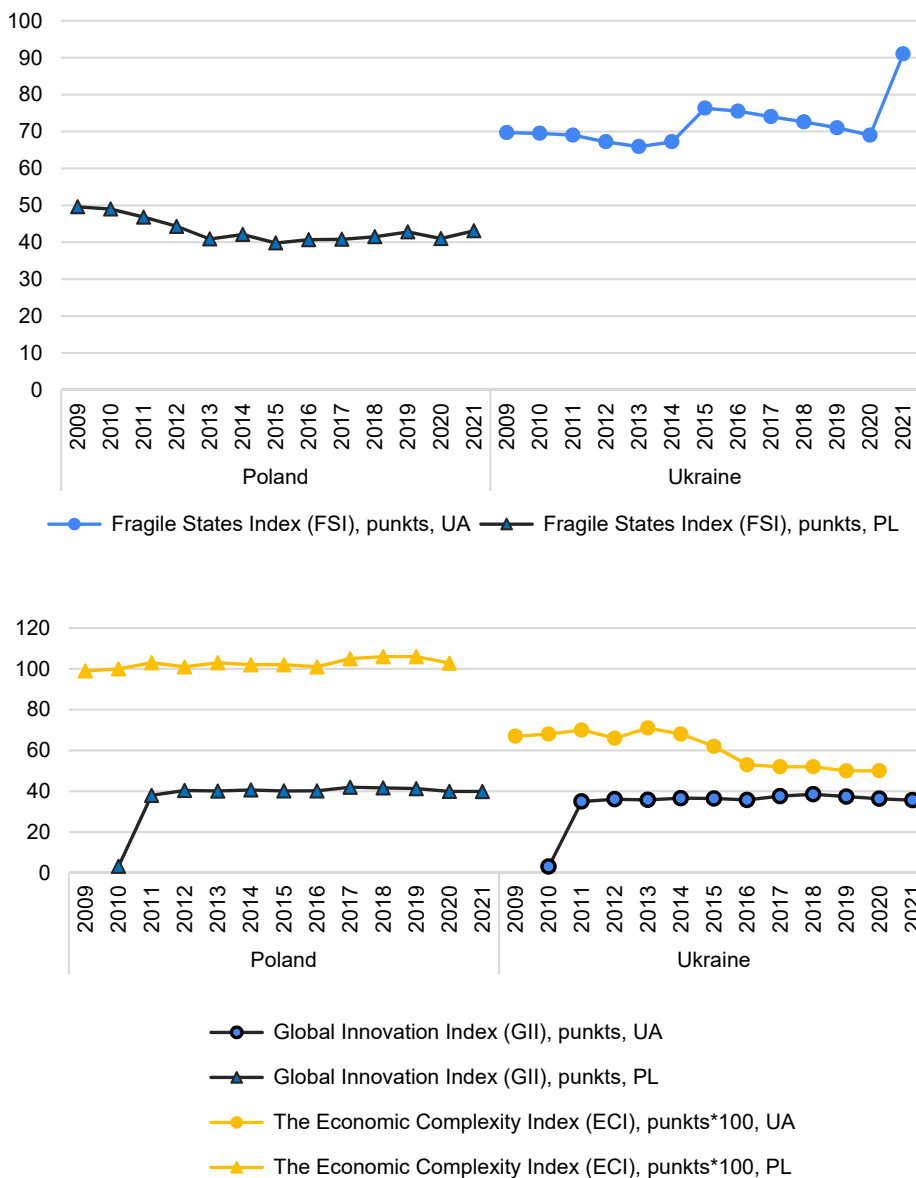
As can be seen in Figure 2, the values of the GII for Poland are always slightly better than for Ukraine. In 2010, when the measurement started, the index for both countries was very low: 3.06 for Ukraine and 3.28 for Poland. But the next year, it jumped to 35.01 for Ukraine, and 38.02 for Poland. Over the next decade, there was little improvement with both countries reaching the highest value in 2018: 38.52 for Ukraine and 41.7 for Poland. As a result of the Covid-19 pandemic and political instability, the GII for Ukraine decreased to 35.6 in 2021. It should be noted that despite a relatively small change in the index value, Ukraine dropped in the world ranking from 60<sup>th</sup> place in 2011 to 49<sup>th</sup> place in 2021. In the same period, Poland rose from 43<sup>rd</sup> to 40<sup>th</sup> place, also registering a fall in the index value to 39.9.

The third indicator in this group, the Economic Complexity Index (ECI), is a “measure of an economy's capacity, which can be inferred from data connecting locations to the activities that are present in them. It has been shown to predict important macroeconomic outcome, including economic growth (...) [It is calculated using] the Product Complexity Index, (PCI), which is a measure of the complexity required to produce a product or engage in an economic activity and is correlated with the spatial concentration of economic activities”<sup>29</sup>.

---

<sup>28</sup> <https://www.globalinnovationindex.org/about-gii#framework>

<sup>29</sup> <https://oec.world/en/resources/methods#eci-intuitively>



**Figure 2.** A comparison of indicators of sustainable economic development in Poland and Ukraine, 2009–2021<sup>30</sup>

Source: based on data published on the official websites listed in the footnotes<sup>31</sup>.

<sup>30</sup> The available data series for the ECI does not include 2021.

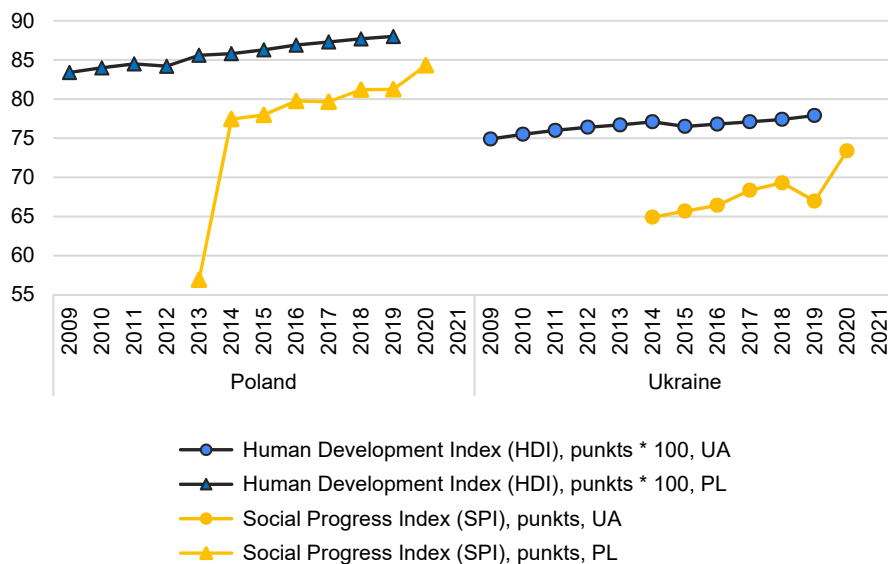
<sup>31</sup> <https://fragilestatesindex.org/>; <https://www.globalinnovationindex.org/Home>; <https://oec.world/en/rankings/eci/hs6/hs96?tab=table>



As shown in Figure 2, for Poland the ECI increased from 0.99 in 2009 to 1.028 in 2020, that is, by 0.039 points, with little effect on the country’s position in the world ranking (from 23<sup>rd</sup> in 2009 to 24<sup>th</sup> in 2020). In same period the value of the index for Ukraine decreased from 0.67 to 0.50, which caused a fall in the ranking from 34<sup>th</sup> to 41<sup>st</sup> place.

The last group of indices in Table 1 includes two social indicators: The Human Development Index (HDI) and Social Progress Index (SPI). The HDI is the most widely used composite measure of average achievement in three dimensions of human development: a long and healthy life, access to education and a decent standard of living. It is a standard tool for general comparisons of the standard of living of different countries and regions and is published as annually part of the UN development program.

As can be seen in Figure 3, both Poland and Ukraine registered a slight increase in the value of the HDI between 2009 and 2019: it rose from 0.834 to 0.88 for Poland (+0.046), but the country’s position in the world ranking (35<sup>th</sup>) remained unchanged. In the case of Ukraine, despite a bigger rise from 0.749 to 0.779 (+0.03), the country actually dropped 2 places in the ranking, from 72<sup>nd</sup> to 74<sup>th</sup> place.



**Figure 3.** A comparison of social indicators for Poland and Ukraine, 2009–2020<sup>32</sup>

Source: based on data published on the official websites listed in the footnotes<sup>33</sup>.

<sup>32</sup> The available data series for the HDI does not include 2020 for Poland and Ukraine; for the SPI does not include 2009-2012 for Poland and 2009-2013 for Ukraine.

<sup>33</sup> <https://ourworldindata.org/human-development-index>; <https://www.socialprogress.org/>

The second indicator in this category is the Social Progress Index (SPI), published by Social Progress Imperative a US-based nonprofit. SPI measures how well countries meet the social and environmental needs of their citizens. It currently uses 60 indicators to measure the social performance of 169 countries<sup>34</sup>. The indicators are grouped into 12 components, which represent three main dimensions of social progress: basic human needs, foundations of well-being and opportunities.

As in the case of the previous indicators, the situation of Poland in terms of the SPI was better than that of Ukraine. Between 2013, when the index was launched, and 2020, the index for Poland rose from 56.92 to 84.32, (+27.4), or almost 1.5 times. It makes no sense to consider the position in the ranking, because the list of participating countries has changed significantly over the years. As for Ukraine, which was included in the ranking in 2014, the initial value of the index was 64.91, which grew to 73.38 in 2020, that is by 13%.

In summary, not only were values of the indices for Poland higher throughout the reference period than those for Ukraine, but their rate of growth was also greater in most cases. The subsequent parts of the study focus on the factors that are directly related to respondent participation in statistical surveys.

#### **4. Basic facts about the latest census rounds in Poland and in Ukraine**

Many countries, including Great Britain, Bulgaria, Hungary, Greece, Italy, Lithuania, Ireland, Poland, Portugal, Estonia, Romania, Czech Republic used to conduct traditional censuses using paper forms and direct interviews. In recent years, NSIs have been increasingly relying on information from administrative registers. The latest censuses in Austria, Denmark, Sweden, Finland and Norway were based entirely on data from state registers; Belgium, Spain, Slovenia, Luxembourg and Latvia combined register data with conventional methods.

Since joining the EU in 2004, Poland has conducted two censuses – in 2011 and 2021. After 2011, Statistics Poland decided to abandon the use of paper forms in favour of the CAPI mode, which contributed to a significant reduction of census costs. The cost of a Ukrainian census in 2023 was estimated to be about USD 243 million USD in September 2021<sup>35</sup>. In comparison, the cost of the 2021 census in Poland, where the number of respondents is comparable to that in Ukraine, was just a third of that amount (about USD 80 million). Also, a lot has been done over the years to modernize the IT infrastructure supporting the collection and processing of statistical information in all departments of Statistics Poland. For example, during the last census data were

---

<sup>34</sup> 2022 Social Progress Index Executive Summary

<sup>35</sup> <https://forbes.ua/news/vtroe-dorozhe-chem-v-polshepochemu-perepis-naseleniya-v-ukraine-stoit-73-mlrd-grn-i-pri-chem-zdes-apple-21092021-2461>

collected by the CAWI method, i.e. using a respondent-friendly online form, which was also accessible via a smartphone app. In addition, the census was preceded and accompanied by a nationwide advertising campaign designed to boost respondent participation in self-enumeration: census ads were visible in public transport, on television, on billboards, in hospitals, schools, etc. Much of the information was obtained from administrative registers even before the start of the census. All these measures contributed to a significant reduction in the census budget and made it possible to conduct it and process the data in a timely manner.

The 2001 census in Ukraine was conducted in a traditional way, using paper forms with a questionnaire containing only 19 questions. In the trial census of 2019, in addition to direct interviews, respondents could provide data via an online form, and information from state registers was also used. For the purpose of the trial census in 2019 and the planned census in 2023, the Ukrainian government approved a questionnaire containing 50 questions (Table 2). The Polish census questionnaire included more questions about gender equality in the family, religion, the use of energy-saving technologies and questions about family connections and health status.

**Table 2.** The number of questions in different categories of the census questionnaire used in Poland and Ukraine

Main categories of census questions	Poland, 2021	Ukraine, 2019
demographic characteristics (age, gender, place of birth, etc.)	7	4
ethnic origin, language, religion, citizenship	9	5
education	1	4
employment	14	4
migration activity	7	8
living conditions	26	24
family connections	5	-
health characteristics, disability status	4	1
<i>Total</i>	73	50

Source: based on data published by Statistics Poland<sup>36</sup> and State Statistics Service of Ukraine<sup>37</sup>.

<sup>36</sup> [https://spis.gov.pl/wp-content/uploads/2021/03/NSP2021\\_Wytyczne-do-samospisu\\_20210311\\_jezyk-polski.pdf](https://spis.gov.pl/wp-content/uploads/2021/03/NSP2021_Wytyczne-do-samospisu_20210311_jezyk-polski.pdf)

<sup>37</sup> <http://www.ukrcensus.gov.ua/>

The budget of the 2001 census amounted to UAH 194.2 million (USD 36.1 million). By the start of 2022, UAH 416 million (USD 14.9 million USD) had already been allocated for the preparation of the census in 2023. According to the forecast of the National Academy of Sciences of Ukraine, the total cost of the 2023 census is expected to be about UAH 6 billion (about USD 214.4 million at the exchange rate from the beginning of 2022)<sup>38</sup>.

## 5. Respondent participation in Ukrainian census, 2019

In order to conduct the trial census in 2019 and the actual census in 2023, it was necessary to adopt new legislation in accordance with new standards and the country's pro-European development. In July 2022, amendments to certain laws of Ukraine regarding state statistical activities were approved. The new legislation amends the laws on state statistics and the law on the all-Ukrainian population census. Some of the key provisions include:

- the requirement of conducting a general population census at least once every 10 years;
- the use of information from administrative registers, in compliance with the requirements of Ukrainian legislation regarding personal data protection;
- the possibility of completing the census form online;
- new requirements for temporary census personnel to guarantee confidentiality of personal information;
- protection of respondents' rights.

The legal changes were made considering international agreements and obligations of Ukraine and are based on the fundamental principles of the UN and EU regarding official statistics, in particular the European Statistics Code of Practice, with the aim of harmonizing the Ukrainian statistical system with European norms and standards.

The trial census held in December of 2019 consisted of the following stages:

- 1) respondents filled out the questionnaire on their own (in response to a census letter) using a special online form;
- 2) those who had not responded online were interviewed by census enumerators equipped with tablets;
- 3) enumerators, together with instructor-controllers, conducted a selective control round of dwellings in order to check the quality of the work of the enumerators and coverage of the trial census.

---

<sup>38</sup> <https://fakty.com.ua/ua/ukraine/suspilstvo/20211130-perepys-naselennya-ukrayiny-u-2023-roczy-yak-organizovuvatymut-ta-skilky-koshtuvatyme/>

According to the State Statistics Service of Ukraine, there were 14882 thousand households in Ukraine. This means that the trial census covered only 0.1% of the country's households. Actually, during the planning stages, the coverage goals in terms of the number of respondents were not set. The main goal was to check the capacity of statistical services, the quality of tablets used by enumerators, to identify weaknesses in the questionnaire and develop proposals for improving the implementation of the survey on the places.

During the trial census in Ukraine, the level of population participation was quite high and amounted to 11.9 out of 15.7 thousand households, or 75.8%. However, among them, only 0.1 thousand households (0.6%) took part in the observation in an online format, without visiting census points, but using a special application with an electronic form in their gadgets. Information about the settings of this application was not widespread enough among young and middle-aged people. At the same time, information from administrative registers was not used for technical reasons. The budget of the trial census was 1.4 billion UAH or 54.3 million USD. In the next part of the study, the authors set the task of selecting such factors that can further contribute to increasing the level of respondents' participation in the Ukrainian national census, a special role in this case belongs to the possibility of online participation based on the use of modern digital technologies.

## **6. Factor analysis of international indicators for Poland and Ukraine using Principal Components Analysis**

Before analyzing which indicators to select as the main components of the measure of respondent participation during in statistical surveys, it was clear that for some indices data in the time series were missing. In particular, the EPFI and EGDI are calculated by the World Economic Forum once every two years; the GII for both countries only became available from 2010, the ECI data for 2021 are not available yet. There were also no HDI data for 2020-2021 at the time when this article was prepared. The SPI index was launched in 2014 and for 2021 were also unavailable at the time when this article was prepared. For this reason, missing data were imputed based using a regression equation of the general trend and checking its reliability by calculating the coefficient of determination. The years with missing data and imputed values used in further analysis are presented in Table 3.

The original data series with values imputed for missing years were used in Principal Components Analysis to select factors that have the biggest effect on respondent participation in statistical surveys in Poland and Ukraine. The analysis was performed using the SPSS software. Figure 4 shows the correlation matrix for the selection of factors using PCA. The first column in each table contains values for 2009-

2021. The variable numbers in Figures 4 and 5 should be reduced by one to denote the index numbers shown in Table 3.

**Table 3.** Data imputation for missing values of indices for Poland and Ukraine, 2009–2021

Index	Years with missing data	Poland		Ukraine	
		Regression equation, coefficient of determination	Imputed value	Regression equation, coefficient of determination	Imputed value
1.EPFI	2009	$y=0.1489\exp(0.1745x)$ $R^2=0.9240$	0.1773	$y=0.1511\exp(0.1527x)$ $R^2=0.8844$	0.1760
	2011		0.2513		0.2389
	2013		0.3563		0.3242
	2015		0.5051		0.4400
	2017		0.7161		0.5972
	2019		1.0151		0.8105
	2021		1.4391		1.1000
2. EGDI	2009	$y=0.0238x+0.4873$ $R^2=0.9753$	0.5111	$y=0.0175x+0.4644$ $R^2=0.8798$	0.4819
	2011		0.5587		0.5169
	2013		0.6063		0.5519
	2015		0.6539		0.5869
	2017		0.7015		0.6219
	2019		0.7491		0.6569
	2021		0.7967		0.6919
3. FSI	-	-	-	-	-
4. GII	2009	$y=-0.548x^2+8.6568x+10.719$ ; $R^2=0.5782$	10.72	$y=-0.4891x^2+7.739x+9.8714$ ; $R^2=0.5727$	9.87
5. ECI(%)	2021	$y=0.4573x+99.594$ ; $R^2=0.5459$	105.5	$y=-2.1115x+74.483$ ; $R^2=0.7909$	47.0
6. HDI(%)	2020	$y=0.4755x+82.938$ $R^2=0.9771$	88.6	$y=-1.1609x+75.425$ $R^2=0.6402$	77.4
	2021		89.1		77.5
7. SPI	2009	$y=-0.7574x^2+9.4411x+54.146$ $R^2=0.7597$	19.01	$y=0.1x^3-1.3107x^2+4.9871x+60.687$ $R^2=0.7848$	12.09
	2010		32.23		30.69
	2011		43.95		44.51
	2012		54.15		54.27
	2013		56.92		60.69
	2021		77.77		78.14

Sources: authors' calculations.

As can be seen from the correlation matrices, for both countries, there is a strong correlation between the first two indices, classified as indicators of electronic document circulation, namely the E-Participation Framework Index (EPFI) and the UN Global E-Government Development Index (EGDI). In the case of Poland, a rather strong correlation can also be observed in the case of the last two indices included in the group of social indicators, namely the Human Development Index (HDI) and the Social Progress Index (SPI).

**Poland**

**Correlation Matrix**

		VAR00002	VAR00003	VAR00004	VAR00005	VAR00006	VAR00007	VAR00008
Correlation	VAR00002	1,000	,933	-,523	,478	,713	,954	,756
	VAR00003	,933	1,000	-,647	,570	,771	,987	,851
	VAR00004	-,523	-,647	1,000	-,837	-,520	-,700	-,901
	VAR00005	,478	,570	-,837	1,000	,647	,620	,813
	VAR00006	,713	,771	-,520	,647	1,000	,779	,679
	VAR00007	,954	,987	-,700	,620	,779	1,000	,874
	VAR00008	,756	,851	-,901	,813	,679	,874	1,000

**Ukraine**

**Correlation Matrix**

		VAR00002	VAR00003	VAR00004	VAR00005	VAR00006	VAR00007	VAR00008
Correlation	VAR00002	1,000	,879	,704	,464	-,919	,766	,753
	VAR00003	,879	1,000	,537	,529	-,904	,754	,772
	VAR00004	,704	,537	1,000	,165	-,627	,392	,411
	VAR00005	,464	,529	,165	1,000	-,414	,805	,852
	VAR00006	-,919	-,904	-,627	-,414	1,000	-,625	-,657
	VAR00007	,766	,754	,392	,805	-,625	1,000	,970
	VAR00008	,753	,772	,411	,852	-,657	,970	1,000

**Figure 4.** Correlation matrices of the factor analysis based on PCA for Poland and Ukraine

Sources: calculated in the SPSS software.

The total variance distribution table is presented in Figure 5. In the case of Poland, only one common component is proposed, which accounts for 78.157% of the total variance of the investigated indices. In the case of Ukraine, the PCA method revealed

two components that explain 89.186% of the total variance. Thus, the residual variance of the influence of the other indices is 21.843% for Poland, and 10.814% for Ukraine.

### Poland

#### Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5,471	78,157	78,157	5,471	78,157	78,157
2	,904	12,911	91,069			
3	,413	5,901	96,970			
4	,119	1,701	98,672			
5	,051	,730	99,402			
6	,040	,566	99,968			
7	,002	,032	100,000			

Extraction Method: Principal Component Analysis.

### Ukraine

#### Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5,052	72,171	72,171	5,052	72,171	72,171
2	1,191	17,014	89,186	1,191	17,014	89,186
3	,433	6,183	95,369			
4	,177	2,527	97,896			
5	,095	1,355	99,251			
6	,037	,527	99,778			
7	,016	,222	100,000			

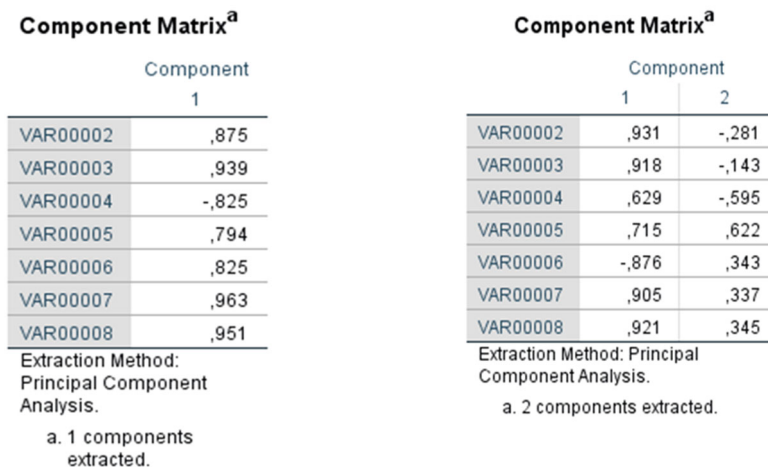
Extraction Method: Principal Component Analysis.

**Figure 5.** Distribution of the total variance into main components for Poland and Ukraine

Sources: calculated in the SPSS software.

Having evaluated the component matrix included in Figure 6, the indices to be included in the main component were provided. In the case of Poland, almost all the selected indices were assigned to the main component, which is why only one group was created for all indices. The only exception is the fourth indicator, the Global Innovation Index (GII), for which the total correlation coefficient is 0.794.





**Figure 6.** Matrices of the main components for Poland and Ukraine

Sources: *calculated in the SPSS software.*

In the case of Ukraine, the software recommends separating two main components. The first main component includes the first, seventh, second and sixth indices: the E-Participation Framework Index (EPFI), the Social Progress Index (SPI), the UN Global E-Government Development Index (EGDI) and the Human Development Index (HDI), their total correlation coefficients are 0.931, 0.921, 0.918 and 0.905, respectively.

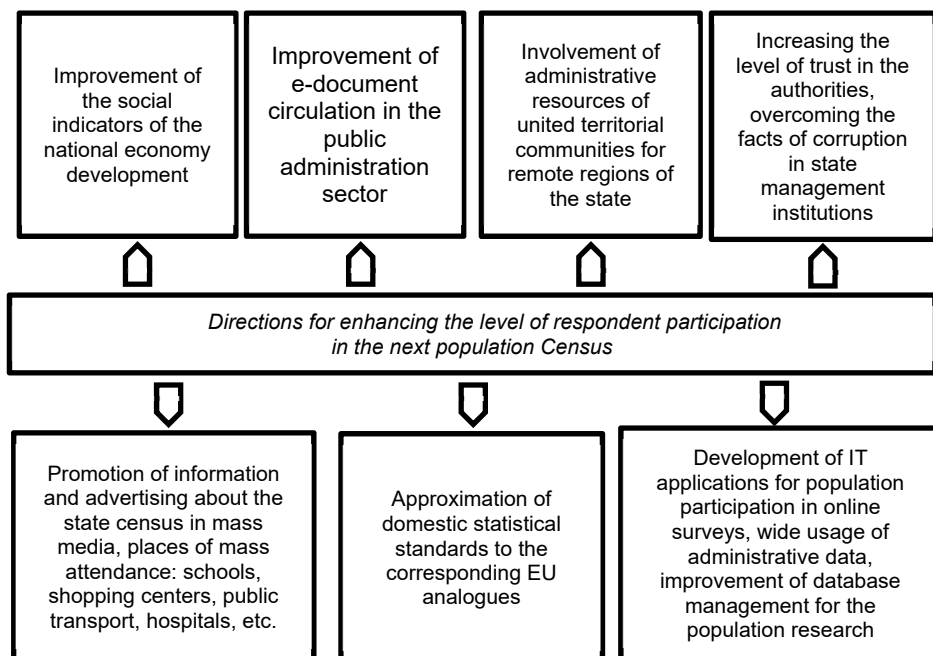
It can thus be concluded that the best way of stimulating respondent participation in statistical surveys and censuses is to improve electronic document circulation, which is measured by the first two indices, namely the E-Participation Framework Index and the UN Global E-Government Development Index. The analysis has also shown that the two social indicators are strongly correlated, which could have an effect on respondent participation. Poland’s experience in organizing and conducting the census suggests that the introduction of digital technologies in the public administration sector plays a significant role. After all, the indicators of electronic document circulation were included in the group of main components in the analysis of international indices, in the case of both countries. The degree of citizens’ e-participation in various events organized by the state depends on the volume and quality of online services, the level of telecommunication infrastructure and the availability of online services for citizens (Figure 7).

As demonstrated by the statistical analysis above, the various social indicators are also important in this context: the quality of life, access to education, medical services, indicators of life expectancy and health, income and expenses of households. It is obvious that the socio-humanitarian component, together with the return of peace, will encourage Ukrainian migrants to return home.

Given that about 30% of the population in most regions live in rural areas, it is important to ensure that the census can be conducted in remote regions, where enumerators could find it difficult to visit respondents (pensioners, disabled people, people with special needs). In such cases, cooperation with representatives of united territorial communities is necessary to make sure that the census can be conducted by employing workers from local communities or social services instead of professional enumerators.

Another important measure is nationwide census promotion in the mass media. It is important to create a positive attitude to the participation in the census and to emphasize its importance for the country's European aspirations. It will also be necessary to adopt uniform classifiers and terminology in accordance with EU standards when processing collected data. For this reason, European statistical standards should be promoted in Ukrainian statistics.

To encourage the use of smartphones, the mobile app with the census form should be adjusted to run on phones with different operating systems. Such an app is likely to appeal to young people and middle-aged people who would not need to visit the territorial district at their place of residence. It could also be used by those who have temporarily gone abroad or migrated in connection with hostility actions, but they are citizens of the state, who should participate in the census.



**Figure 7.** Directions for improving the level of respondent participation in official statistical surveys

Sources: results of performed authors' research.

## 7. Conclusions

The authors have analyzed a number of international indices that are relevant for respondent participation in statistical surveys, and particularly in the context of the next population census. Three groups of indices were identified: indicators of electronic document circulation, indicators of sustainable economic development, and social indicators. Considering each of these indices, the situation of Poland is significantly better compared to that of Ukraine, where the last national census was conducted only once in 2001.

A comparative analysis of census questionnaires used in Poland in 2021 and in Ukraine in 2019 revealed that the Polish census form was not only longer (73 vs 50 questions), but also included some aspects that were absent from the Ukrainian questionnaire, e.g. a section about family ties in the household.

As regards respondent participation, a very low percentage of young respondents self-enumerated online, probably because of insufficient information about how to use the web application.

The factor analysis of the indices for Poland and Ukraine was carried out using Principal Component Analysis. The matrix analysis of the main components for Poland showed that almost all the selected indices had an impact on the successful implementation of the National Census. In the case of Ukraine, the main components include the E-Participation Framework Index (EPFI), the Social Progress Index (SPI), the UN Global E-Government Development Index (EGDI) and the Human Development Index (HDI).

The authors have proposed a number of recommendations regarding the main areas that should be improved to ensure a successful administration of the next census in Ukraine. The most important of these measures include the digitization of the public administration sector, the use of smartphone apps and census promotion in the mass media.

## Acknowledgements

The project entitled „Census administration in Ukraine: insight into the Polish experience in the context of international indicators analysis” is financed by the Polish National Science Centre DEC-2013/11/B/HS4/01472.

## References

- About the Sustainable Development Goals of Ukraine for the period until 2030, (2019). Decree of the President of Ukraine dated September 30, URL: <https://zakon.rada.gov.ua/laws/show/722/2019#Text> [in Ukrainian].
- Dunteman, G., (1989). Principal components analysis. Newbury Park. URL: <https://www.worldcat.org/title/principal-components-analysis-dunteman/oclc/1042959545>
- Dygaszewicz, J., (2020). Transition from traditional census to combined and registers-based census. *Statistical Journal of the IAOS*. 36(1), pp. 165–175
- Fichman, M., (2003). Multiple imputation for missing data: Making the most of what you know. Graduate School of Industrial Administration Carnegie-Mellon University. URL: [https://kithub.cmu.edu/articles/journal\\_contribution/Multiple\\_Imputation\\_for\\_Missing\\_Data\\_Making\\_the\\_Most\\_of\\_What\\_you\\_Know/6707012/1](https://kithub.cmu.edu/articles/journal_contribution/Multiple_Imputation_for_Missing_Data_Making_the_Most_of_What_you_Know/6707012/1).
- Gołata, E., Dehnel, G., (2021). Credibility of disability estimates from the 2011 population census in Poland. *Statistics in Transition new series*, vol. 22, 2, pp. 41–65. URL: <https://sit.stat.gov.pl/Article/208>.
- Guidelines for the self-census in the People's National Census in Poland, (2021). URL: [https://spis.gov.pl/wp-content/uploads/2021/03/NSP2021\\_Wytyczne-do-samospisu\\_20210311\\_jezyk-polski.pdf](https://spis.gov.pl/wp-content/uploads/2021/03/NSP2021_Wytyczne-do-samospisu_20210311_jezyk-polski.pdf) [in Polish].
- Da Costa, J. N., Bielecka, E., Calka, B., (2017). Uncertainty quantification of the global rural-urban mapping project over Polish census data. 10th International Conference on Environmental Engineering, ICEE 2017 enviro.2017.221.
- Kolomiets, O., (2022). How will we gather you all?. Why will not all refugees return to Ukraine? *Ekonomichna Pravda*, 30.05.2022. URL: <https://www.epravda.com.ua/publications/2022/05/30/687530/> [in Ukrainian].
- Kravchenko, V. P., Kravchenko, N. V., (2015). The current demographic situation in Ukraine and prospects for its development. *Visnik sotsialno-ekonomichnih doslidzhen: zbiinik nauk. prats; za red. M.I. Zveryakova ta in.* Odesa. Odeskiy natsionalniy ekonomichniy universitet, vol. 3, 58, pp. 236–240 [in Ukrainian].
- Krywult-Albańska, M., (2012). Spis powszechny jako źródło informacji o ludności [National Census as a source of data on population]. *Studia Socjologiczne*, (4), pp. 87–107 [in Polish].

- Libanova, E., (2019), Labour migration from Ukraine: key features, drivers and impact. *Economics and Sociology*, vol. 12(1), pp. 313–328. URL: [https://www.economics-sociology.eu/?657,en\\_labour-migration-from-ukraine-key-features-drivers-and-impact](https://www.economics-sociology.eu/?657,en_labour-migration-from-ukraine-key-features-drivers-and-impact).
- Human development of the regions of Ukraine: analysis and forecast, (2007). Edited by E. Libanova. Institute of Demography and Social Research of the National Academy of Sciences of Ukraine. URL: [http://irbis-nbuv.gov.ua/cgi-bin/ua/elib.exe?Z21ID=&I21DBN=UKRLIB&P21DBN=UKRLIB&S21STN=1&S21REF=10&S21FMT=online\\_book&C21COM=S&S21CNR=20&S21P01=0&S21P02=0&S21P03=FF=&S21STR=ukr0005357](http://irbis-nbuv.gov.ua/cgi-bin/ua/elib.exe?Z21ID=&I21DBN=UKRLIB&P21DBN=UKRLIB&S21STN=1&S21REF=10&S21FMT=online_book&C21COM=S&S21CNR=20&S21P01=0&S21P02=0&S21P03=FF=&S21STR=ukr0005357) [in Ukrainian].
- Malish, N. A., (2016). Demographic aspects of socio-humanitarian development. URL: [http://www.akademy.gov.ua/ej8/doc\\_pdf/malysh.pdf](http://www.akademy.gov.ua/ej8/doc_pdf/malysh.pdf) [in Ukrainian].
- Malynovska, O., (2016). Migration in Ukraine: facts and figures. Kyiv: Mizhnarodna orhanizatsiia z Mihratsii. Predstavnytstvo v Ukraini. URL: [https://iom.org.ua/sites/default/files/ff\\_ukr\\_21\\_10\\_press.pdf](https://iom.org.ua/sites/default/files/ff_ukr_21_10_press.pdf) [in Ukrainian].
- Marciniak, G., (2014). Modern approach to censuses in the case of Poland-Advantages and constraints. *Statistical Journal of the IAOS*. 30(1), pp. 29–34
- Melnik, S. I., (2014). Demographic situation in Ukraine: state, main problems and ways to solve them. Ukraine: aspects of work. *Ukrayina: aspekti pratsi*, 4, pp. 22–26 [in Ukrainian].
- Mokin B. I., Mokin O. B., (2015). Methodology and organization of scientific research: textbook / 2nd edition, amended and supplemented. Vinnitsya: VNTU, 317. [in Ukrainian].
- National Population and Housing Census 2021. Research methodology and organization, (2022). Statistics Poland. Warsaw. URL: <https://stat.gov.pl/en/national-census/national-population-and-housing-census-2021/national-population-and-housing-census-2021/national-population-and-housing-census-2021-research-methodology-and-organization,3,1.html>.
- On approval of forms of Census documentation for conducting a Trial Population Census in 2019 and instructions for filling them out, (2019). Order of the State Statistics Service of Ukraine dated November 19, No. 372. URL: <https://ukrstat.gov.ua/> [in Ukrainian].
- On the adoption as a basis of the draft Law of Ukraine, (2022). On amendments to some laws of Ukraine regarding state statistical activity. Resolution of the Verkhovna

Rada of Ukraine dated July 1, URL: <https://zakon.rada.gov.ua/laws/show/2349-20#Text> [in Ukrainian].

On the approval of the Procedure and conditions for providing subventions from the state budget to local budgets for the implementation of projects within the framework of the Program for the Reconstruction of Ukraine. Resolution of the Cabinet of Ministers of Ukraine dated 15.12.2021, No. 1324. URL: <https://www.minregion.gov.ua/napryamki-diyalnosti/international-cooperation/spivpraczya-z-mizhnarodnymy-finansovymy-organizacziyamy/yevropejskyj-investycziyjnyj-bank/programa-z-vidnovlennya-ukrayiny/postanova-kabinetu-ministriv-ukrayiny-vid-15-12-2021-%e2%84%96-1324-pro-zatverdzhennya-poryadku-ta-umov-nadannya-subvencziyi-z-derzhavnogo-byudzhetu-miszczevym-byudzhetam-na-realizacziyu-proektiv/> [in Ukrainian].

Osaulenko, O. , Bulatova, O., Zakharova, O., Reznikova N., (2021). The problem of statistical assessment of the potential for the development of regional integration processes. *Statistics in Transition new series*, vol. 22, 4, pp. 121–138. URL: <https://sit.stat.gov.pl/Article/243>.

Preliminary results of the National Population and Housing Census 2021. URL: <https://stat.gov.pl/en/national-census/national-population-and-housing-census-2021> [in Polish].

Program for the Reconstruction of Ukraine. Ministry of Development of Communities and Territories of Ukraine. URL: <https://www.minregion.gov.ua/napryamki-diyalnosti/international-cooperation/spivpraczya-z-mizhnarodnymy-finansovymy-organizacziyamy/yevropejskyj-investycziyjnyj-bank/programa-z-vidnovlennya-ukrayiny/> [in Ukrainian].

Schwab K., Zahidi S., (2020). The Global Competitiveness Report. How Countries are Performing on the Road to Recovery. World Economic Forum. URL: [https://www3.weforum.org/docs/WEF\\_TheGlobalCompetitivenessReport2020.pdf](https://www3.weforum.org/docs/WEF_TheGlobalCompetitivenessReport2020.pdf).

Szymkowiak, M., Wilak, K., (2021). Repeated weighting in mixed-mode censuses. *Economics and Business Review*, 7(1), pp. 26–46.

The Cabinet of Ministers proposes to conduct a population Census once every 10 years, (2021). Word and deed. Analytical portal from February 17, URL: <https://www.slovoidilo.ua/2021/02/17/novyna/suspilstvo/kabmin-proponuye-provodyty-perepys-naselennya-raz-10-rokiv> [in Ukrainian].

UN, (2008a). Principles and Recommendations for Population and Housing Censuses Revision 2, New York: United Nations.

- UN, (2008b). The measurement of Disability Recommendations for the 2010 Round of Censuses.
- UN, (2015a). Principles and recommendations for population and housing censuses - Revision 3. In Statistical Papers Series M., No. 67/Rev. 3. New York: United Nations.
- Wisla, R., Chugaievska, S., Nowosad, A., Turanli, U., (2020). Structural changes in the Polish and Ukrainian economies against the background of the Central and Eastern European countries. Economic transformation in Poland and Ukraine. National and regional perspectives / Edited by Rafal Wisla and Andrzej Nowosad. London and New York: Routledge Taylor & Francis Group, pp. 41–56. URL: <https://www.routledge.com/Economic-Transformation-in-Poland-and-Ukraine-National-and-Regional-Perspectives/Wisla-Nowosad/p/book/9780367484934#sup>.
- Zayukov, I. V., (2011). Components of the current demographic crisis in Ukraine and their impact on labor potential. *Ukrayina: aspekti pratsi*, 4, pp. 41–46 [in Ukrainian].





# Estimating the probability of leaving unemployment for older people in Poland using survival models with censored data

Wioletta Grzenda<sup>1</sup>

## Abstract

Current demographic changes require greater participation of people aged 50 or older in the labour market. Previous research shows that the chances of returning to employment decrease with the length of the unemployment period. In the case of older people who have not reached the statutory retirement age, these chances also depend on the time they have left to retirement. Our study aims to assess the probability of leaving unemployment for people aged 50-71 based on their characteristics and the length of the unemployment period. We use data from the Labour Force Survey for 2019–2020. The key factors determining employment status are identified using the proportional hazard model. We take these factors into account and use the direct adjusted survival curve to show how the probability of returning to work in Poland changes as people age. Due to the fact that not many people take up employment around their retirement age, an in-depth evaluation of the accuracy of predictions obtained via the models is crucial to assess the results. Hence, in this paper, a time-dependent ROC curve is used. Our results indicate that the key factor that influences the return to work after an unemployment period in the case of older people in Poland is whether they reached the age of 60. Other factors that proved important in this context are the sex and the education level of older people.

**Key words:** employment, older workers, proportional hazard model, time-dependent ROC curve.

## 1. Introduction

Given the demographic changes taking place in Poland, resulting in a decrease in labour supply and an increase in the old-age dependency ratio, it is necessary to boost the participation of older people in the labour force. Understanding at the individual level the factors that favour and limit leaving unemployment for people aged 50 or more may significantly increase the effectiveness of activities related to their return to employment.

---

<sup>1</sup> Institute of Statistics and Demography, Collegium of Economic Analysis, SGH Warsaw School of Economics, Poland. E-mail: [wgrzend@sgh.waw.pl](mailto:wgrzend@sgh.waw.pl). ORCID: <https://orcid.org/0000-0002-2226-4563>.



The job-finding chances vary depending on the duration of unemployment (Sheldon, 2020; Jarosch, 2021). According to Charni (2022), the probability of returning to employment after an unemployment period decreases as workers get older. Moreover, for people aged 50 or older who do not yet have pension rights, the time remaining to obtain them is also important (Hairault et al., 2010). This study aims to assess the probability of returning to work for people aged 50–71, based on their characteristics and the length of unemployment.

The data for our study were obtained from the Labour Force Survey (LFS) from 2019–2020. Considering that not only the chances of finding a job may change over time, but also their determinants, our study focused on short-term unemployment and medium-term unemployment. Therefore, only those people who had been unemployed for a maximum of 12 months at the time of the survey were considered. The Cox regression model (Cox, 1972; Cox and Oakes, 1984) was used to identify the factors determining return to work for people aged 50–71. Then, based on this model, with the direct adjusted survival curve (Chang et al., 1982; Gail and Byar, 1986; Zhang et al., 2007) it was shown how the probability of returning to work for these people changes over time.

The use of survival models makes it possible to obtain time-dependent results, but it is associated with the analysis of censored data, which poses many challenges to the development of this class of models. The ROC curve and the area under the ROC curve (AUC) are common tools for assessing the discrimination ability of a regression model with a binary dependent variable. These methods are also used in the case of survival models, where the time-dependent ROC curve is considered (Heagerty et al., 2000; Heagerty and Zheng, 2005; Kamarudin et al., 2017). This curve is determined for various time points, which makes it possible to evaluate predictive accuracy at specific times (Guo and Jang, 2017).

The low employment rate among people aged 50 or more in Poland (Eurostat, 2022) reflects a small number of people around retirement age taking up employment. Therefore, in this paper, we focus on the predictive accuracy of the obtained results. Given that the obtained prediction changes over time, the time-dependent ROC curve was used in this study. To the best of our knowledge, this is the first study for Poland in which the probability of leaving unemployment for people aged 50 or more was estimated based on their characteristics, and predictive accuracy evaluation over time was performed.

This paper is structured as follows. The first part presents the labour force behaviour of older people in Poland and the determinants of their employment. A description of the methods used is provided in Section 3, which is followed by a description of the data (Section 4). Section 5 presents the results of our analyses. Section 6 provides a discussion and conclusive remarks.

## 2. The employment of people aged 50 years or older

In studies on employment of people aged 50 years or older, much attention is paid to the causes of ending their occupational careers and their withdrawal from the labour market (Gałęcka-Burdziak and Góra, 2016; Jansen, 2018; Phillipson et al., 2016). Resigning from work at an older age often results in permanent withdrawal from the labour market. This is reflected, *inter alia*, in the differences between the statutory retirement age and the effective retirement age in Poland. In 2018, the effective retirement age in Poland was 60.6 for women and 62.8 for men (OECD, 2019), while the statutory retirement age was 60 and 65, respectively. In the same year, the employment rate for people aged from 50 to 74 years in Poland amounted to 40.2%, and by 2020 it increased by only 0.1% (Eurostat, 2022). Compared to the European average (27 countries of the European Union) of 47% in 2020, this puts Poland among the European countries where employment of older people is very low. Increasing the labour force participation of older people in Poland could, to some extent, limit the effects of population aging.

Many factors may affect the duration of the working life of people aged 50 years or older. Some factors may push these people out of the labour market, while others encourage them to stay employed. Among them, we distinguish characteristics of individuals, such as sex, age, education, skills, work experience, or place of residence, and factors directly related to the work, such as working hours, the time needed to reach a workplace, occupation, and many others. Based on Eurostat data, it can be concluded that in the case of older people, sex is the key factor affecting employment rates, in addition to age (Eurostat, 2022). Women in OECD countries, and in particular in Poland (OECD, 2019), often withdraw from the labour market much earlier than men. According to Blackburn et al. (2016), gender inequality in employment is reflected not only in the length of working life but also after retirement.

According to Rutledge et al. (2017), gender differences also occur in the chances of finding a job in the years preceding retirement. The authors indicate that the range of occupations in which employment can be found changes with age, and their availability depends on sex and education. The narrowing of the number of occupations mainly takes place at the age of 50 for less-educated men, and at the age of 60 for women and better-educated men. However, the authors make it clear that the employment opportunities for better-educated older workers have expanded significantly since the late 1990s. Also, according to Torp (2015), having a greater level of human capital in terms of education and skills may enable older people to stay active and productive for a longer time. However, Bowman et al. (2017) draw attention to the obsolescence of older workers' job skills, which is a significant problem limiting employment opportunities in older age. Moreover, the authors argue that the competitiveness of

older workers in the labour market is not only a function of their knowledge and technical job skills. Consequently, competencies currently desired in the labour market cannot be acquired through investments made by an individual.

According to the results of a study by Charni (2022) based on British panel data, the age of employees has a large impact on their chances of getting back to work after unemployment, in addition to human capital characteristics and economic incentives. Older jobseekers are longer unemployed than younger jobseekers (Bowman et al., 2017). According to Charni (2022), the time it takes for older people to return to employment after an unemployment spell would be shorter if they were treated in the same way as younger people. This result indicates the need to combat age discrimination also in the workplace and when looking for employees. According to Fleischmann et al. (2015), the treatment of older workers is influenced not only by the characteristics of the organization but most of all by the local labour market. A significant drop in labour supply may significantly impact how older workers are perceived by employers in Poland.

### 3. Methods

In the survival analysis, the most popular basic function is the survival function:

$$S(t) = P(T > t), \quad (1)$$

where  $T$  is the variable describing the time until the event occurs. The Kaplan-Meier method (Kaplan and Meier, 1958) or the Nelson-Aalen method (Aalen, 1978; Nelson, 1972) are most often used to estimate this function. An alternative to these methods is the Cox regression model approach (Cox, 1972; Cox and Oakes, 1984):

$$h(t) = h_0(t) \exp(\mathbf{x}'\boldsymbol{\beta}). \quad (2)$$

In the proportional hazard model, the formula for the survival function is given as follows:

$$S(t) = [S_0(t)]^{\exp(\mathbf{x}'\boldsymbol{\beta})}, \quad (3)$$

where  $\boldsymbol{\beta}$  is a vector of estimated model parameters, and  $S_0(t)$  is a baseline survival function corresponding to a baseline hazard  $h_0(t)$ . The baseline survival function  $S_0$  can be written using the cumulative hazard function  $H_0$  as follows:

$$S_0(t) = \exp(-H_0(t)), \quad (4)$$

where  $H_0(t) = \int_0^t h_0(u) du$ ,  $t \geq 0$ . The estimator of the survival function  $S(t)$  can be written in the following form:

$$\hat{S}(t) = [\hat{S}_0(t)]^{\exp(\mathbf{x}'\hat{\boldsymbol{\beta}})}, \quad (5)$$

where  $\widehat{\boldsymbol{\beta}}$  denotes the estimator of the parameter vector  $\boldsymbol{\beta}$ , and  $\widehat{S}_0$  is an estimator of a baseline survival function, which is given by the following formula:

$$\widehat{S}_0(t) = \prod_{u|t_{(u)} < t} \left( 1 - \frac{d_u}{\sum_{l \in R(t_{(u)})} \exp(\mathbf{x}'_l \widehat{\boldsymbol{\beta}})} \right), \tag{6}$$

where  $d_u$ ,  $u = 1, 2, \dots, m$  is the number of observations, for which the event occurred at time  $t_{(u)}$ ,  $u = 1, 2, \dots, m$ , and  $R(t_{(u)})$ ,  $u = 1, 2, \dots, m$ , denotes a hazard set. The hazard set includes all individuals for which the survival or censoring time is greater than  $t_{(u)}$ .

Let  $j$  denote an individual belonging to the  $k$ -th group, then the observed values for this individual can be described by  $\{t_{kj}, v_{kj}, \mathbf{x}_{kj}\}$ ,  $k = 1, 2, \dots, K, j = 1, 2, \dots, n_k$ , where  $t_{kj}$  is the observed time,  $v_{kj} = 0$ , when censoring occurs and  $v_{kj} = 1$ , otherwise, and  $\mathbf{x}_{kj}$  denotes a covariates vector. Then, the survival function at time point  $t$ , for an individual from the  $k$ -th group, with values of variables  $\mathbf{x}$ , is given by the formula (Chang et al., 1982; Gail and Byar, 1986; Zhang et al., 2007):

$$\widehat{S}_k(t; \mathbf{x}) = \exp\{-\widehat{H}_{0k}(t) \exp(\mathbf{x}' \widehat{\boldsymbol{\beta}})\}. \tag{7}$$

Then, the general formula for the direct adjusted survival function estimator has the following form:

$$\widehat{S}_k(t) = \frac{1}{n} \sum_{l=1}^n \exp\{-\widehat{H}_{0k}(t) \exp(\mathbf{x}'_l \widehat{\boldsymbol{\beta}})\}, \tag{8}$$

where  $n = \sum_{k=1}^K n_k$ .

The Cox regression results can also be used to obtain estimates of a time-dependent sensitivity and a time-dependent specificity, and in consequence to obtain estimates of the time-dependent ROC curve (Heagerty et al., 2000; Heagerty and Zheng, 2005).

Let  $T$  denote a variable describing the time until the event occurs, and  $Z_i = \mathbf{x}'_i \boldsymbol{\beta}$  for  $i$ -th individual ( $i = 1, 2, \dots, n$ ). Moreover, let  $D_i(t)$  denote status of  $i$ -th individual at time  $t$  defined as follows:

$$D_i(t) = I(T \leq t). \tag{9}$$

Then, for a given cut-off point  $c$ , the time-dependent sensitivity ( $Se$ ) and the time-dependent specificity ( $Sp$ ) can be defined as follows (Heagerty et al., 2000; Heagerty and Zheng, 2005; Kamarudin et al., 2017):

$$Se(c, t) = P(Z_i > c | D_i(t) = 1), \tag{10}$$

$$Sp(c, t) = P(Z_i \leq c | D_i(t) = 0). \tag{11}$$

Let TPR (True Positive Rate) be given by the formula  $TPR(c, t) = Se(c, t)$ , and FPR (False Positive Rate) will be calculated as follows:  $FPR(c, t) = 1 - TNR(c, t) = 1 - Sp(c, t)$ , where TNR denotes True Negative Rate. Thus, the time-dependent ROC

curve (ROC) can be determined for any time points  $t$  and for varying cut-off points  $c$  as follows:

$$ROC(t) = \{(FPR(c, t), TPR(c, t)): c \in \mathbf{R}\}. \quad (12)$$

Then, the time-dependent AUC is defined as follows:

$$AUC(t) = \int_{-\infty}^{+\infty} Se(c, t)d[1 - Sp(c, t)]. \quad (13)$$

#### 4. Data

The estimation of the probability of returning to work for people aged 50-71 was made based on the data from the LFS for Poland for 2019 and 2020. The study included people who during the first wave of the survey conducted in 2019 answered both the question "During the week in question, Monday through Sunday, did you do any work for at least one hour that generated income or earnings, or did you assist on an unpaid basis in a family business?", and the question "Did you have a job in the surveyed week, but did not perform it temporarily?" with "no". Short-term and medium-term unemployment was studied, therefore only those who had stopped working no more than 12 months before the first wave of the survey was performed in 2019 were included in the study. As many as 619 people were identified, and only 5.33% of them took up employment in the analysed period. For the survival analysis, it was assumed that an event occurred for these people.

The time was calculated in months from the moment of leaving the last job until the beginning of work or until the end of the observation period, i.e. the moment of the last wave of the survey in which the respondent participated. The observation time ranged from 1 month to 27 months. When constructing the variable describing the age of a respondent, the different retirement ages for women and men in Poland were considered. This variable was defined in such a way as to determine those who have not yet retired. The set of other individual characteristics included in the study is presented in Table 1. Due to a small share of people from the central macroregion and difficulties in estimating appropriate coefficients for the needs of the analysis, the central macroregion was combined with the southwestern macroregion.

**Table 1:** Sample characteristics

Variable	Categories	Percent
Sex	Men	46.20
	Women	53.80
Age	50-54 years old	8.72
	55-60 years old	22.29
	61-65 years old	48.47
	66-71 years old	20.52

**Table 1:** Sample characteristics (cont.)

Variable	Categories	Percent
Marital status	Single	22.78
	Married	77.22
Educational level	Higher	15.02
	Post-secondary or secondary professional or secondary general	31.34
	Basic vocational	33.93
	Primary school	19.71
Last job type	Self-employed	12.92
	Other	87.08
Employment sector - last job	Public	29.89
	Private	70.11
Place of residence	City 100,000 residents or more	33.12
	City from 20,000 to 100,000 residents	21.16
	City under 20,000 residents	15.83
	Rural areas	29.89
Macroregion	Southern (Regions: małopolskie, śląskie)	13.41
	North-Western (Regions: wielkopolskie, zachodniopomorskie, lubuskie)	19.06
	South-Western (Regions: dolnośląskie, opolskie)	11.79
	Northern (Regions: kujawsko-pomorskie, warmińsko-mazurskie, pomorskie)	17.61
	Central (Regions: łódzkie, świętokrzyskie)	8.08
	Eastern (Regions: lubelskie, podkarpackie, podlaskie)	17.13
	Mazovian Province (Regions: warszawski stołeczny, mazowiecki regionalny)	12.92
Were you looking for a job in the last 4 weeks?	Yes	9.21
	No, because I already have a job and I am waiting for it to start	1.78
	No	89.01

Source: Own calculations; data from Labour Force Survey 2019 and 2020, Poland.

## 5. Results

In the first stage of this research, the factors influencing the employment of people aged 50-71 were identified with the proportional hazard model. The use of this model required prior verification of the assumed hazard proportionality. For this purpose, time-dependent variables were incorporated into the model. We found out that this assumption is fulfilled for all considered variables. The estimates of the proportional hazards model parameters are presented in Table 2.

It was found that men had over two times greater hazard of taking up employment than women. However, the key factors that affected the employment of respondents were their age and educational level. People aged 50–54 had an 8.37 times greater hazard of taking up employment than people aged 66–71. A similar result was obtained for people aged 55–60, they had a 7.28 times greater hazard of taking up employment compared to the oldest people. In the case of people aged 61–65, no statistically significant differences were observed compared to the oldest people. People with a high level of education had a 12.91 times greater hazard of taking up employment than people with primary education. People with post-secondary or secondary professional or secondary general education had a 10.92 times greater hazard of taking up employment compared to people with primary education. People with vocational education had a 6.9 times greater hazard of taking up employment compared to the least-educated people. People who had worked in the public sector in their last job had a 67.5% lower hazard of taking up employment, compared to people who had worked in the private sector in their last job. In the case of the variable describing the type of place of residence, only one level turned out to be statistically significant. People who lived in cities of under 20,000 residents had a 2.6 times greater hazard of taking up employment than people who had lived in rural areas. The remaining levels of the variable describing the type of place of the residence turned out to be statistically insignificant. Moreover, in the model, we included two control variables: a variable describing the macroregion of residence and information if the respondent was looking for a job in the last 4 weeks.

**Table 2:** Estimated parameters, standard error, *p*-value, and hazard ratio – results from the proportional hazards model

Covariate	Parameter estimate	Standard error	<i>p</i> -value	Hazard ratio
<i>Sex (ref. Women)</i>				
Men	0.8130	0.4180	0.0518	2.255
<i>Age (ref. 66–71 years)</i>				
50–54 years old	2.1245	0.9778	0.0298	8.369
55–60 years old	1.9856	0.9297	0.0327	7.283
61–65 years old	0.1663	0.9735	0.8644	1.181
<i>Educational level (ref. Primary school)</i>				
Higher	2.5581	1.0248	0.0126	12.912
Post-secondary or secondary professional or secondary general	2.3904	0.8854	0.0069	10.917
Basic vocational	1.9312	0.8603	0.0248	6.898

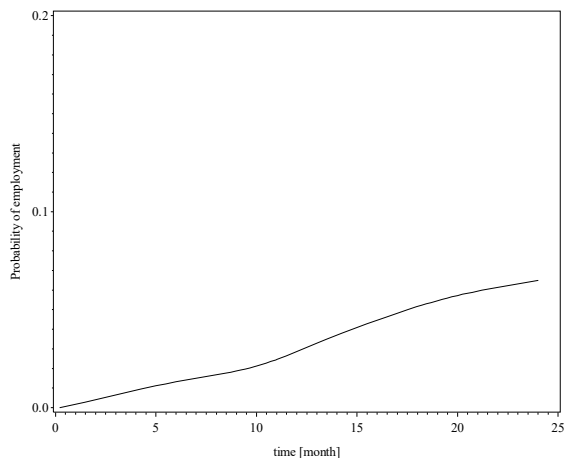


**Table 2:** Estimated parameters, standard error, *p*-value, and hazard ratio – results from the proportional hazards model (cont.)

Covariate	Parameter estimate	Standard error	p-value	Hazard ratio
Employment sector - last job ( <i>ref. Private</i> )				
Public	-1.1225	0.6573	0.0877	0.325
Place of residence ( <i>ref. Rural areas</i> )				
City 100,000 residents and more	0.0511	0.5318	0.9235	1.052
City from 20,000 to 100,000 residents	-0.2432	0.5141	0.6361	0.784
City under 20,000 residents	0.9582	0.5518	0.0824	2.607
Macroregion - Mazovian Province ( <i>ref. Regions: warszawski stołeczny, mazowiecki regionalny</i> )				
Southern Macroregion (małopolskie, śląskie)	1.3110	0.7512	0.0809	3.710
North-Western Macroregion (wielkopolskie, zachodniopomorskie, lubuskie)	1.1785	0.7363	0.1095	3.250
South-Western Macroregion (dolnośląskie, opolskie) and centralny (łódzkie, świętokrzyskie)	-1.4186	1.0018	0.1567	0.242
Northern Macroregion (kujawsko-pomorskie, warmińsko-mazurskie, pomorskie)	1.7406	0.7157	0.0150	5.701
Eastern Macroregion (lubelskie, podkarpackie, podlaskie)	1.3310	0.7512	0.0764	3.785
Were you looking for a job in the last 4 weeks? ( <i>ref. No</i> )				
Yes	1.9524	0.4496	<.0001	7.045
No, because I already have a job and I am waiting for it to start	3.6336	0.7051	<.0001	37.849

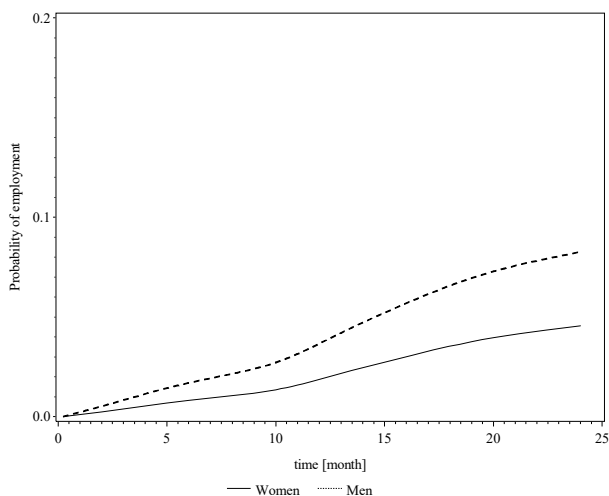
Source: Own calculations; data from Labour Force Survey 2019 and 2020, Poland.

Based on Cox regression with the direct adjusted survival curve, it was shown how the probability of returning to work for respondents changes over survey time. It was found that this probability increased very slowly, and it did not even exceed the value of 0.1 (Figure 1).



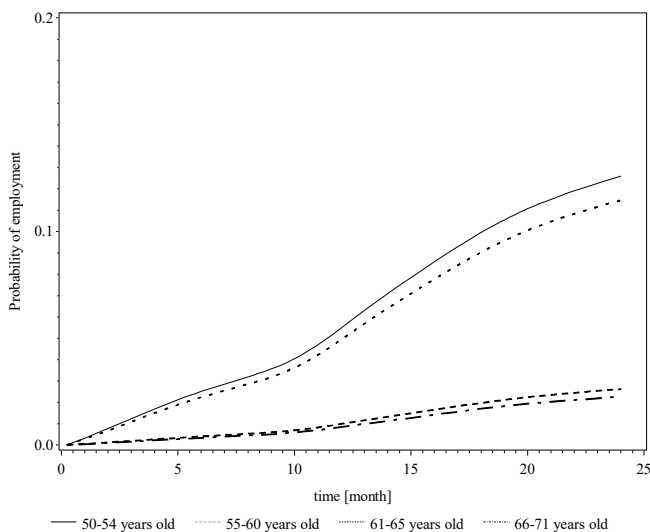
**Figure 1:** The probability of leaving unemployment for all individuals determined with the direct adjusted survival curves

Due to the different retirement ages for women and men in Poland, the direct adjusted survival curve was estimated in the groups determined by sex. It was found that after about 10 months the probability of taking up employment increased for both women and men, but in the case of men this increase was higher (Figure 2). In the next stage of this research, the direct adjusted survival curve was estimated in the groups defined by the age of the respondents. It was found that in the analysed period the probability of leaving unemployment for people aged 50–54 is very similar to the probability of leaving unemployment for people aged 55–60.



**Figure 2:** The probability of leaving unemployment for women and men determined with the direct adjusted survival curves

The similarity was observed for people aged 61–65 and people aged 66–71 too (Figure 3). People aged 50–54 and 55–60 had a greater probability of taking up employment than older people. This probability started to increase after 10 months but ultimately did not exceed the value of 0.2. In the case of people aged 61–65 and 66–70, the increase in the probability of taking up employment was very small – the probability of taking up employment remained at a very low level throughout the entire period under study.



**Figure 3:** The probability of taking up employment by age determined with the direct adjusted survival curves

Among the surveyed respondents, only 5.33% took up employment in the period under study. Due to such a small percentage of events, the evaluation of predictive accuracy was performed. For this purpose, time-dependent ROC curves were used. Based on the obtained results, it can be concluded that our prediction is better than the random one over the whole analysed period (Figure 4). Moreover, the predictive accuracy increases with time, starting from the 5th month. Additionally, Figures 5–8 show the shape of ROC curves at selected time points (after 6, 12, 18, and 24 months). The shapes of the time-dependent ROC curves confirm the high quality of the previously presented probability estimates for leaving unemployment for the surveyed respondents aged 50 years or older. The AUC at 6 months was 0.764, at 12 months 0.836, at 18 months 0.841, and at 24 months it reached 0.874.

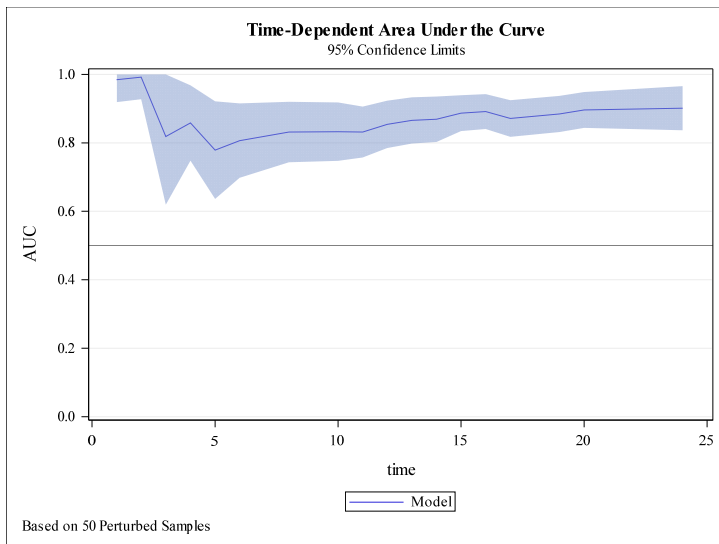


Figure 4: The time-dependent area under the ROC curve and the 95% confidence limits

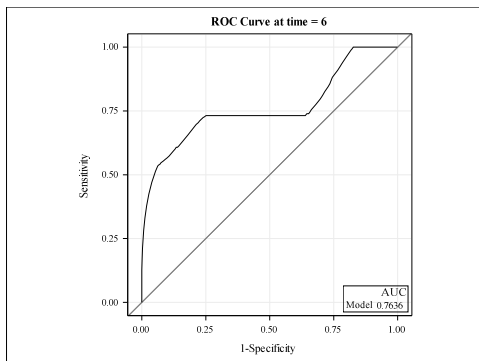


Figure 5: ROC curve at 6 months

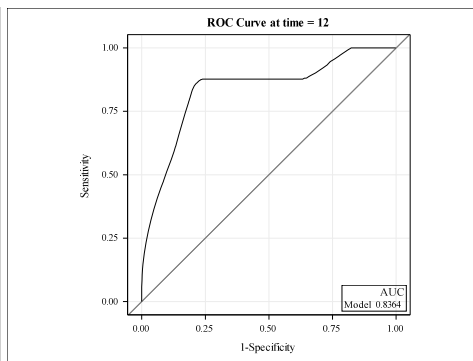
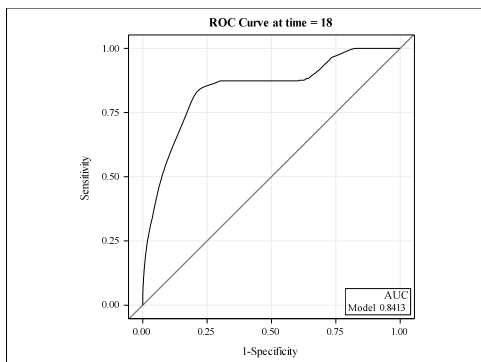
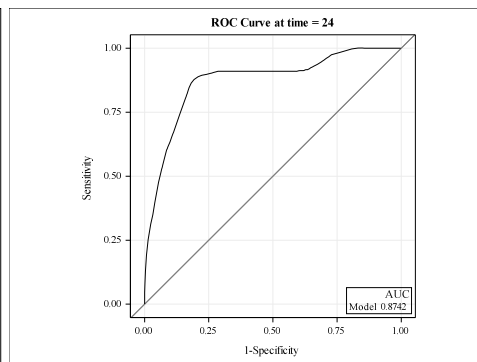


Figure 6: ROC curve at 12 months



**Figure 7:** ROC curve at 18 months



**Figure 8:** ROC curve at 24 months

## 6. Discussion and conclusion

A further increase in the demographic dependency ratio may pose a threat to the stability of the pension system in Poland. Therefore, it is necessary to take measures that would effectively encourage the elderly to stay longer in the labour market, even after reaching retirement age. This study raises the issue of returning to work after a break in employment by people aged 50 years or older in Poland. Moreover, prolonged unemployment of people in the pre-retirement age may cause permanent withdrawal from the labour market.

In the first stage of the research, the factors determining return to work for people aged 50–71 were identified with the proportional hazards model. It was found that the age of the respondents had a large impact on taking up employment. People who had not yet reached the statutory retirement age for women in Poland (60 years) had over seven times greater hazard of taking up employment than those aged 66–71. However, no statistically significant differences were found between people aged 61–65 and people aged 66–71. This result is similar to the results of previous studies based on data from other countries indicating the major importance of age in the case of older people in returning to employment after an unemployment spell (Bowman et al., 2017; Charni, 2022). Moreover, we showed the large impact of reaching the statutory retirement age on taking up employment after a break in employment for the surveyed respondents.

The other key factors that affected the return to employment of people aged 50–71 after an unemployment spell were sex and educational level. We revealed that men had over two times more chances of employment than women. This result is in line with previous research results for other countries (Blackburn et al., 2016; Rutledge et al., 2017). However, taking into account the result obtained for Poland and the 16.6% difference (Eurostat, 2022) in the value of employment rate for women and men aged 50 to 74 in 2020 in Poland this problem seems to particularly affect the Polish labour

market. Moreover, our research indicates that also the level of education was relevant for leaving unemployment for people aged 50–71. The previous study based on US data (Rutledge et al., 2017) also indicates such a relationship. In addition, according to Batyra et al. (2019), low-skilled senior workers demonstrate a higher probability of not only unemployment but also of early withdrawal from the labour market.

The results of the second part of our research revealed that the probability of leaving unemployment for respondents aged 50–71 in the period under consideration increased very slowly over time, and ultimately did not even exceed the value of 0.1. Looking at our result as well as earlier findings of Charni (2022) it can be concluded that only a few unemployed people who are around retirement age return to work after an unemployment spell. Moreover, we showed how the probability of taking up employment changes over time depending on sex. It was found that after 10 months the probability of taking up employment increases for both women and men, but in the case of men this increase is slightly higher. The analysis of the probability of taking up employment depending on age confirmed the earlier conclusions that the key time point followed by a decline in the chances of returning to work is reached at the age of 61. In addition, we have shown the relationship between returning to work and the period during which unemployment benefits were paid, which is usually 6 or 12 months. After about 10 months, the probability of taking up employment increased, albeit slightly, for all analysed groups of respondents.

The time-dependent ROC curve analyses allow the conclusion that obtained predictive accuracy increases with time and is at a satisfactory level during the entire period under observation (after 24 months the AUC reached the value of 0.874).

Based on our results, it can be concluded that it is crucial to take actions that will promote the continuity of employment of people in the pre-retirement age in Poland because any return to employment after an unemployment spell is very unlikely for people aged 50 years or older.

## References

- Aalen, O. O., (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, 6, pp. 701–726.
- Batyra, A., Pierrard, O., de la Croix, D. and Sneessens, H. R., (2019). Structural changes in the labor market and the rise of early retirement in France and Germany. *German Economic Review*, 20(4), pp. e38–e69.
- Blackburn, R. M., Jarman, J. and Racko, G., (2016). Understanding gender inequality in employment and retirement. *Contemporary Social Science*, 11(2-3), pp. 238–252.

- Bowman, D., McGann, M., Kimberley, H. and Biggs, S., (2017). 'Rusty, invisible and threatening': ageing, capital and employability. *Work, employment and society*, 31(3), pp. 465–482.
- Chang, I. M., Gelman, R. and Pagano, M., (1982). Corrected group prognostic curves and summary statistics. *Journal of chronic diseases*, 35(8), 669–674.
- Charni, K., (2022). Do employment opportunities decrease for older workers? *Applied Economics*, 54(8), pp. 937–958.
- Cox, D. R., (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), pp. 187–202.
- Cox, D.R., Oakes, D., (1984). *Analysis of Survival Data*. London – New York: Chapman and Hall.
- Eurostat, (2022). *Employment rates by sex, age and citizenship (%)*.  
[https://ec.europa.eu/eurostat/databrowser/view/lfsa\\_ergan/default/table?lang=en](https://ec.europa.eu/eurostat/databrowser/view/lfsa_ergan/default/table?lang=en)
- Hairault, J. O., Sopraseuth, T. and Langot, F., (2010). Distance to retirement and older workers' employment: The case for delaying the retirement age. *Journal of the European Economic association*, 8(5), pp. 1034–1076.
- Heagerty, P. J., Lumley, T. and Pepe, M. S., (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2), pp. 337–344.
- Heagerty, P. J., Zheng, Y., (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1), pp. 92–105.
- Gail, M. H., Byar, D. P., (1986). Variance calculations for direct adjusted survival curves, with applications to testing for no treatment effect. *Biometrical Journal*, 28(5), pp. 587–599.
- Guo, C., So, Y. and Jang, W., (2017). *Evaluating predictive accuracy of survival models with PROC PHREG. Paper SAS462-2017*, Cary, NC: SAS Institute.
- Fleischmann, M., Koster, F. and Schippers, J., (2015). Nothing ventured, nothing gained! How and under which conditions employers provide employability-enhancing practices to their older workers. *The International Journal of Human Resource Management*, 26(22), pp. 2908–2925.
- Gałecka-Burdziak, E., Góra, M., (2016). The impact of easy and early access to old-age benefits on exits from the labour market: a macro-micro analysis. *IZA Journal of European Labor Studies*, 5(1), pp. 1–18.

- Jansen, A., (2018). Work–retirement cultures: a further piece of the puzzle to explain differences in the labour market participation of older people in Europe? *Ageing & Society*, 38(8), pp. 1527–1555.
- Jarosch, G., (2021). Searching for job security and the consequences of job loss. *Working Paper No. 28481, National Bureau of Economic Research*.  
[https://www.nber.org/system/files/working\\_papers/w28481/w28481.pdf](https://www.nber.org/system/files/working_papers/w28481/w28481.pdf)
- Kamarudin, A. N., Cox, T. and Kolamunnage-Dona, R., (2017). Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC medical research methodology*, 17(1), pp. 1–19.
- Kaplan, E.L., Meier, P., (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, pp. 457–481.
- Nelson, W. (1972). Theory and application of hazard plotting for censored survival data. *Biometrics*, 14, pp. 945–966.
- OECD, (2019). *Pensions at a Glance 2019: OECD and G20 Indicators*. Paris: OECD Publishing.
- Sheldon, G., (2020). Unemployment in Switzerland in the wake of the Covid-19 pandemic: an intertemporal perspective. *Swiss Journal of Economics and Statistics*, 156(1), pp. 1–9.
- Phillipson, C., Vickerstaff, S. and Lain, D., (2016). Achieving fuller working lives: Labour market and policy issues in the United Kingdom. *Australian Journal of Social Issues*, 51(2), pp. 187–203.
- Rutledge, M. S., Sass, S. A. and Ramos-Mercado, J. D., (2017). How does occupational access for older workers differ by education? *Journal of Labor Research*, 38(3), pp. 283–305.
- Torp, C., (Ed.), (2015). *Challenges of aging: Pensions, retirement and generational justice*. New York: Palgrave Macmillan.
- Zhang, X., Loberiza, F. R., Klein, J. P. and Zhang, M. J., (2007). A SAS macro for estimation of direct adjusted survival curves based on a stratified Cox regression model. *Computer methods and programs in biomedicine*, 88(2), pp. 95–101.



# Bayesian estimation of a geometric distribution using informative priors based on a Type-I censoring scheme

Nadeem Akhtar<sup>1</sup>, Sajjad Ahamad Khan<sup>2</sup>, Muhammad Amin<sup>3</sup>,  
Akbar Ali Khan<sup>4</sup>, Amjad Ali<sup>5</sup>, Sadaf Manzoor<sup>6</sup>

## Abstract

In this paper, the geometric distribution parameter is estimated under a type-I censoring scheme by means of the Bayesian estimation approach. The Beta and Kumaraswamy informative priors, as well as five loss functions are used for this purpose. Expressions of Bayes estimators and Bayes risks are derived under the Squared Error Loss Function (SELF), the Quadratic Loss Function (QLF), the Precautionary Loss Function (PLF), the Simple Asymmetric Precautionary Loss Function (SAPLF), and the DeGroot Loss Function (DLF) using the two aforementioned priors. The prior densities are obtained through prior predictive distributions. Simulation studies are carried out to make comparisons using Bayes risks. Finally, a real-life data example is used to verify the model's efficiency.

**Key words:** prior distribution, posterior distribution, geometric distribution, beta distribution, Kumaraswamy distribution.

## 1. Introduction

Type-I censored sampling is helpful for lifetime research. Type-I censored data are used when the last few observations in a series are missing or suppressed due to lack of time or because the experimenter cannot wait for the last observation due to time restriction. Shi and Yan (2010) produced type-I censored empirical Bayes estimates of the two-parameter exponential distribution. Saleem et al. (2010) used type-I censored data to explore the power function mixture distribution. Tahir et al. (2016) study the Bayesian analysis of a three-component mixture of exponential distributions. Khan et al. (2016) designed and compared different loss functions for estimating scale parameter of log-normal distribution under type-I censoring schemes. Yanuar et al. (2019) used Bayesian estimation tool to estimate the scale parameter of Weibull distribution. Kour et al. (2020) employed Bayesian and E-Bayesian techniques to estimate exponential-Lomax distribution's parameters. Abbas et al. (2020) used Bayesian inference to estimate the parameters of Gumbel type-II distribution under censored sample scenario. Long (2021) estimated Rayleigh distribution's parameters

<sup>1</sup>Department of Statistics Islamia College Peshawar, Pakistan. E-mail: nadeemscholar@icp.edu.pk.  
ORCID: <https://orcid.org/0000-0002-2169-5185>.

<sup>2</sup>Department of Statistics Islamia College Peshawar, Pakistan. E-mail: sajjadkhan@icp.edu.pk.  
ORCID: <https://orcid.org/0000-0001-6630-7222>.

<sup>3</sup>Nuclear Institute for Food and Agriculture (NIFA), Peshawar, Pakistan. E-mail: aminkanju@gmail.com.

<sup>4</sup>Government Postgraduate College, Pakistan.

<sup>5</sup>Department of Statistics Islamia College Peshawar, Pakistan.

<sup>6</sup>Department of Statistics Islamia College Peshawar, Pakistan.



using double Type-I hybrid censored data. Most of the studies considered continuous life testing models. This study explores the estimation issue for the parameter of a discrete Geometric life testing model in the Bayesian paradigm using different loss functions and priors information under the type-1 censoring sampling scheme. The posterior distributions are derived using Beta and Kumarasawmy priors. In addition, the posterior Bayes estimators (BE) and Bayes risks (BR) are derived using the Squared Error Loss Function (SELF), Quadratic Loss Function (QLF), Precautionary Loss Function (PLF), Simple Asymmetric Precautionary Loss Function (SAPLF), and DeGroot Loss Function (DLF). A simulation study is carried out with different sample sizes and different parametric settings in order to make numerical comparisons. A real data set is also used to validate the simulation findings.

## 2. Posterior distributions under different priors

The probability density function of the geometric distribution for a random variable  $X$  is given below.

$$f(x) = \omega(1 - \omega)^x, \quad x = 0, 1, \dots \quad (1)$$

where  $\omega$  is the parameter of the Geometric distribution. The cumulative distribution function of the geometric distribution for a random variable  $X$  is given by:

$$F_X(x) = 1 - (1 - \omega)^{x+1}, \quad x = 0, 1, \dots \quad (2)$$

The likelihood function of the Geometric distribution under type-I censoring sampling scheme can be written as:

$$L(\omega) \propto \prod_{i=1}^r \omega(1 - \omega)^{x_i} [(1 - \omega)^{t+1}]^{n-r} \quad (3)$$

It is assumed that the parameter follows Beta distribution, i.e.  $\omega \sim \text{Beta}(f_1, f_2)$  with hyper-parameters  $f_1$  and  $f_2$ , using this prior, the posterior distribution of  $\omega$  given data is:

$$p_{\beta}(\omega | x) = \frac{\omega^{r+f_1-1} (1 - \omega)^{\delta_{\beta}-1}}{B(f_1 + r, \delta_{\beta})}, \quad 0 < \theta < 1 \quad (4)$$

## 3. Elicitation of Hyper-Parameters

The Aslam (2003) approach is used to elicit the informative priors. For this purpose, prior distributions derived under Beta and Kumaraswamy priors are used. We consider two intervals for two unknown hyper parameters of Beta prior, i.e. (0, 2) and (6, 8), with the associated probabilities 0.6001 and 0.100 as an expert's belief about these intervals. The hyper-parameters of Beta prior are obtained as follows.

$$f_1 = 0.77930, \quad f_2 = 1.4340$$

The resultant value of the hyper-parameter of Kumaraswamy prior is  $f_3 = 1.241$ .

### 4. Loss functions

Under the assumed priors, BEs and BRs are determined under the SELF, DLF, QLF, PLF and SAPLF. The expressions are provided in the following tables. Under the assumed priors, BEs and BRs are determined under the SELF, DLF, QLF, PLF and SAPLF. The expressions are provided in the following tables.

**Table 1.** Bayes estimators and Bayes Risks under SELF

Priors	BE	BR
Beta	$\frac{B(r+f_1+1, \delta_\beta)}{B(r+f_1, \delta_\beta)}$	$\frac{B(r+f_1+2, \delta_\beta)}{B(r+f_1, \delta_\beta)} - (BE)^2$
Kumarswamy	$\frac{B(r+2, \delta_{KS})}{B(r+1, \delta_{KS})}$	$\frac{B(r+3, \delta_{KS})}{B(r+1, \delta_{KS})} - (BE)^2$

**Table 2.** Bayes estimators and Bayes Risks under DLF

Priors	BE	BR
Beta	$\frac{B(r+f_1+2, \delta_\beta)}{B(r+f_1+1, \delta_\beta)}$	$1 - \frac{(B(r+f_1+1, \delta_\beta))^2}{B(f_1+r, \delta_\beta)B(r+f_1+2, \delta_\beta)}$
Kumarswamy	$\frac{B(r+3, \delta_{KS})}{B(r+2, \delta_{KS})}$	$1 - \frac{(B(r+2, \delta_{KS}))^2}{B(r+1, \delta_{KS})B(r+3, \delta_{KS})}$

**Table 3.** Bayes estimators and Bayes Risks under QLF

Priors	BE	BR
Beta	$\frac{B(r+f_1-1, \delta_\beta)}{B(r+f_1-2, \delta_\beta)}$	$1 - \frac{(B(r+f_1-1, \delta_\beta))^2}{B(f_1+r, \delta_\beta)B(r+f_1-2, \delta_\beta)}$
Kumarswamy	$\frac{B(r, \delta_{KS})}{B(r-1, \delta_{KS})}$	$1 - \frac{(B(r, \delta_{KS}))^2}{B(r+1, \delta_{KS})B(r-1, \delta_{KS})}$

**Table 4.** Bayes estimators and Bayes Risks under PLF

Priors	BE	BR
Beta	$\sqrt{\frac{B(r+f_1+2, \delta_\beta)}{B(\omega_1+r, \delta_\beta)}}$	$2 \left( \sqrt{\frac{B(r+f_1+2, \delta_\beta)}{B(f_1+r, \delta_\beta)}} - \frac{B(r+f_1+1, \delta_\beta)}{B(f_1+r, \delta_\beta)} \right)$
Kumarswamy	$\sqrt{\frac{B(r+3, \delta_{KS})}{B(r+1, \delta_{KS})}}$	$2 \left( \sqrt{\frac{B(r+3, \delta_{KS})}{B(r+1, \delta_{KS})}} - \frac{B(r+2, \delta_{KS})}{B(r+1, \delta_{KS})} \right)$

**Table 5.** Bayes estimators and Bayes Risks under SAPLF

Priors	BE	BR
Beta	$\sqrt{\frac{B(r+f_1+1, \delta_\beta)}{B(r+f_1-1, \delta_\beta)}}$	$2 \left( \frac{1}{B(f_1+r, \delta_\beta)} \times \sqrt{\frac{B(r+f_1+1, \delta_\beta)}{B(r+f_1-1, \delta_\beta)}} - 1 \right)$
Kumarswamy	$\sqrt{\frac{B(r+2, \delta_{KS})}{B(r, \delta_{KS})}}$	$2 \left( \frac{1}{B(r+1, \delta_{KS})} \times \sqrt{\frac{B(r+2, \delta_{KS})}{B(r, \delta_{KS})}} - 1 \right)$

### 5. Simulations Study

From a lifetime model of geometric distribution, random samples are generated with samples of sizes  $n = 20$  and  $50$  by considering different termination times and different parametric values, simulation process is performed 10,000 times. Based on these samples, BEs and BRs are obtained. The findings of the simulation study are presented in Tables 6 – 15.

**Table 6.** BEs and BRs under SELF for  $T = 5$

Priors	$n$	$\omega = 0.4$		$\omega = 0.5$		$\omega = 0.6$	
		BEs	BRs	Bes	BRs	BEs	BRs
Beta	20	0.398549	0.0048122	0.497784	0.0058672	0.5928	0.0065180
	50	0.391299	0.001964	0.494964	0.0024617	0.595256	0.0027738
Kumaraswamy	20	0.402898	0.0048229	0.502888	0.0058569	0.598677	0.0064739
	50	0.393052	0.0019670	0.497044	0.0024608	0.597672	0.0027669

**Table 7.** BEs and BRs under SELF for  $T = 7$

Priors	$n$	$\omega = 0.4$		$\omega = 0.5$		$\omega = 0.6$	
		BEs	BRs	Bes	BRs	BEs	BRs
Beta	20	0.405604	0.0046816	0.501316	0.0057882	0.594287	0.0064853
	50	0.400007	0.0019197	0.500538	0.0024332	0.59743	0.0027578
Kumaraswamy	20	0.4098	0.0046908	0.506334	0.0057780	0.600126	0.0064418
	50	0.401704	0.0019217	0.502591	0.0024320	0.599834	0.0027508

**Table 8.** BEs and BRs under DLF for  $T = 5$

Priors	$n$	$\omega = 0.4$		$\omega = 0.5$		$\omega = 0.6$	
		BEs	BRs	Bes	BRs	BEs	BRs
Beta	20	0.399297	0.004819	0.497697	0.0058707	0.591013	0.0065107
	50	0.392174	0.001969	0.494613	0.0024616	0.594912	0.0027743
Kumaraswamy	20	0.403651	0.004829	0.502801	0.0058606	0.596875	0.0064675
	50	0.39393	0.001972	0.496694	0.0024607	0.597328	0.0027674

**Table 9.** BEs and BRs under DLF for  $T = 7$

Priors	$n$	$\omega = 0.4$		$\omega = 0.5$		$\omega = 0.6$	
		BEs	BRs	Bes	BRs	BEs	BRs
Beta	20	0.405099	0.0046789	0.501034	0.0057820	0.593663	0.0064793
	50	0.399603	0.0019172	0.500154	0.0024312	0.597708	0.0027578
Kumaraswamy	20	0.409296	0.0046880	0.50605	0.0057718	0.599497	0.0064360
	50	0.401299	0.0019193	0.502206	0.0024301	0.600114	0.0027508

**Table 10.** BEs and BRs under QLF for  $T = 5$

Priors	$n$	$\omega = 0.4$		$\omega = 0.5$		$\omega = 0.6$	
		BEs	BRs	Bes	BRs	BEs	BRs
Beta	20	0.38389	0.032751	0.478039	0.026819	0.570583	0.0217898
	50	0.39330	0.012797	0.4907470	0.0103919	0.588859	0.0082929
Kumaraswamy	20	0.38823	0.032165	0.48322	0.0262729	0.576646	0.0212587
	50	0.39501	0.012705	0.4927920	0.0103066	0.591267	0.0082101

**Table 11.** BEs and BRs under QLF for  $T = 7$

Priors	$n$	$\omega = 0.4$		$\omega = 0.5$		$\omega = 0.6$	
		BEs	BRs	Bes	BRs	BEs	BRs
Beta	20	0.384534	0.0316685	0.478954	0.0264339	0.571199	0.0216775
	50	0.394012	0.0123857	0.491567	0.0102510	0.588864	0.0082620
Kumaraswamy	20	0.388734	0.031120	0.484075	0.025902	0.577245	0.0211505
	50	0.395762	0.0122944	0.493359	0.0101678	0.591263	0.0081797

**Table 12.** BEs and BRs under PLF for  $T = 5$

Priors	$n$	$\omega = 0.4$		$\omega = 0.5$		$\omega = 0.6$	
		BEs	BRs	Bes	BRs	BEs	BRs
Beta	20	0.414857	0.0116282	0.509512	0.0115902	0.600445	0.0109775
	50	0.406027	0.0048758	0.504217	0.0048912	0.599809	0.0046379
Kumaraswamy	20	0.418962	0.0115406	0.514367	0.0114678	0.606064	0.0108136
	50	0.407697	0.0048613	0.506212	0.0048706	0.602139	0.0046097

**Table 13.** BEs and BRs under PLF for  $T = 7$

Priors	$n$	$\omega = 0.4$		$\omega = 0.5$		$\omega = 0.6$	
		BEs	BRs	Bes	BRs	BEs	BRs
Beta	20	0.413877	0.0112833	0.508801	0.0114575	0.600232	0.0109478
	50	0.405998	0.0047290	0.503476	0.0048350	0.600198	0.0046205
Kumaraswamy	20	0.417853	0.0112013	0.513593	0.0113383	0.605831	0.0107851
	50	0.407617	0.0047154	0.505446	0.0048149	0.602523	0.0045925

**Table 14.** BEs and BRs under SAPLF for  $T = 5$

Priors	$n$	$\omega = 0.4$		$\omega = 0.5$		$\omega = 0.6$	
		BEs	BRs	Bes	BRs	BEs	BRs
Beta	20	0.40045	0.031393	0.496323	0.0254192	0.588079	0.0205064
	50	0.402002	0.0125155	0.49928	0.0101448	0.59496	0.0081071
Kumaraswamy	20	0.4048	0.030806	0.501463	0.0248808	0.594022	0.0199894
	50	0.403758	0.0124211	0.501366	0.0100575	0.597385	0.0080231

**Table 15.** BEs and BRs under SAPLF for  $T = 7$

Priors	$n$	$\omega = 0.4$		$\omega = 0.5$		$\omega = 0.6$	
		BEs	BRs	Bes	BRs	BEs	BRs
Beta	20	0.40299	0.0301996	0.496625	0.0250879	0.589734	0.0203528
	50	0.400943	0.0121620	0.499331	0.0100232	0.596173	0.0080540
Kumaraswamy	20	0.407222	0.0296524	0.501705	0.0245629	0.595672	0.0198396
	50	0.402642	0.0120733	0.501392	0.0099378	0.598594	0.0079705

The above numerical results indicate that as the sample size increases, the efficiency of the BE increases as the BR decreases. Also, the increasing test termination time results in smaller BRs and enhanced estimating efficiency. The overall results show that Beta prior is a suitable prior and SELF is an efficient loss function in estimating the unknown parameter of the Geometric life testing model under type-I censoring scheme.

## 6. Applications

In this section, a data set of 137 lung cancer patients' remission duration, given by Krishna and Neha (2017), is analysed with the help of the proposed estimation strategy. The numerical results for this data set are presented in Table 16.

**Table 16.** BEs and BRs for Lung Cancer data

Prior		SELF		DLF	
		T=14	T=19	T=14	T=19
Beta	BEs	0.15662500	0.0934033	0.119137	0.103447
	BRs	0.000178453	0.0000638503	0.00755068	0.00719107
Kumaraswamy	BEs	0.15691800	0.0935679	0.119358	0.103628
	BRs	0.000178718	0.0000639499	0.00753454	0.00717693
Priors		QLF		PLF	
		T=14	T=19	T=14	T=19
Beta	BEs	0.10662500	0.09204240	0.10871200	0.093753800
	BRs	0.00778249	0.00739421	0.000831595	0.000682338
Kumaraswamy	BEs	0.10681700	0.09219790	0.10890400	0.093909000
	BRs	0.00776680	0.00738051	0.000831408	0.000682216
		SAPLF			
		T=14	T=19		
Beta	BEs	0.107866000	0.093060400		
	BRs	0.007753930	0.007370370		
Kumarasawmy	BEs	0.108070000	0.093225000		
	BRs	0.007737310	0.007355840		

## 7. Conclusion

In this study, a Bayesian estimation methodology is derived for estimating the parameter of discrete Geometric life testing model under type-I censoring scheme. Two informative priors, namely Beta and Kumaraswamy, and five loss functions (SELF, QLF, SAPLF, DLF and PLF) are used for this purpose. An extensive simulation study and a real-life data analysis is employed to validate the importance of the proposed strategy. The numerical results reveal that Beta is an appropriate prior and SLEF is a better loss function while analysing discrete Geometric life testing model under type-I censoring scheme. The real-life data analysis cements these findings.

## Acknowledgements

Thanks to anyone for support, the anonymous referee and the associate editor for their constructive comments and suggestions which had greatly improved the earlier version of this manuscript.

## References

- Abbas, K., Hussain, Z., Rashid, N., Ali, A., Taj, M., Khan, S.A., Manzoor, S., Khalil, U. and Khan, D. M., (2020). Bayesian estimation of gumbel type-II distribution under type-II censoring with medical applications. *Computational and Mathematical Methods in Medicine*, pp. 1–11.
- Aslam, M., (2003). An application of prior predictive distribution to elicit the prior density. *Journal of Statistical Theory and applications*, 2(1), pp. 70–83.
- Khan, A. A., Aslam, M., Hussain, Z., & Tahir, M., (2015). Comparison of loss functions for estimating the scale parameter of log-normal distribution using non-informative priors. *Hacetatepe Journal of Mathematics and Statistics*, 45(6), pp. 1831–1845.
- Kour, K., Kumar, P., Anand, P., & Sudan, J. K., (2020). E-Bayesian estimation of Exponential-Lomax distribution under asymmetric loss functions. *International Journal of Applied Mathematics and Statistics, Int. J. Appl. Math. Stat.*, 59(2), pp. 1–26.
- Krishna, H., & Goel, N., (2017). Maximum likelihood and Bayes estimation in randomly censored geometric distribution. *Journal of Probability and Statistics*, 3, pp. 1–12.
- Long, B., (2021). Estimation and prediction for the Rayleigh distribution based on double type-I hybrid censored data. *Communications in Statistics-Simulation and Computation*, pp. 1–15.
- Saleem, M., Aslam, M., & Economou, P., (2010). On the Bayesian analysis of the mixture of power function distribution using the complete and the censored sample. *Journal of Applied Statistics*, 37(1), pp. 25–40.
- Shi, Y., & Yan, W., (2010). The EB estimation of scale-parameter for two parameter exponential distribution under the type-I censoring life test, *Journal of Physical Science*, 4, pp. 25–30.
- Tahir, M., Aslam, M., & Hussain, Z., (2016). On the Bayesian analysis of 3-component mixture of exponential distributions under different loss functions. *Hacetatepe Journal of Mathematics and Statistics*, 45(2), pp. 609–628.
- Yanuar, F., Yozza, H., & Rescha, R. V., (2019). Comparison of two priors in Bayesian estimation for parameter of Weibull distribution. *Science and Technology Indonesia*, 4(2), pp. 82–87.







Statistics Poland and the Polish Statistical Association  
are pleased to announce the organization of the **MET2023**  
Conference on Methodology of Statistical Research which will take place  
from 3 to 5 July 2023 in Warsaw

Confirmed Keynote Speakers: Partha Lahiri and Stefaan Werhulst

This year's Conference provides us with an opportunity to celebrate the 30<sup>th</sup> anniversary of *Statistics in Transition* and its 100<sup>th</sup> issue.

The MET2023 conference will include presentations, speeches, panels and debates arranged in the following thematic blocks:

Mathematical statistics  
Representative method and small area statistics  
Population statistics  
Social statistics  
Economic statistics  
Regional statistics  
Data analysis and classification  
Big Data and statistical data  
Polish statistics in the international arena  
Statistics in public statistics  
History of Polish statistics, statistical methods in historical research  
Communication of statistics and statistical education

For more information on this event look at:

<https://met2023.stat.gov.pl/en>

<https://met2023.stat.gov.pl/en/Program>



## **About the Authors**

**Adebiyi Aliu A.** is a graduate student at the Department of Statistics, Faculty of Science, University of Ibadan, Nigeria. His major research interest is in environmental statistics, biostatistics, statistical methods, and spatial statistics. He has 2 publications to his credit. He is a senior collaborator and instructor at the University of Ibadan Laboratory for Interdisciplinary Statistical Analysis.

**Adesina Oluwaseun A.** is a Lecturer at the Department of Statistics, Faculty of Pure and Applied Science, Ladoko Akintola University of Technology, Ogbomoso, Oyo State, Nigeria. Her research interests are econometrics statistics, Bayesian probability, econometrics time series, demography, general statistics, and health statistics, which focuses on real-life situations, environmental sciences. Dr. Adesina has published research papers in both international and national journals with conferences. She is also a member of editorial board: Asian Journal of Probability and Statistics.

**Akbar Ali Khan** holds the office of Assistant Professor at the Department of Statistics Government Postgraduate College Kohat (Higher Education Department), Khyber Pakhtunkhwa, Pakistan. He is also part of Quaid-i-Azam University Islamabad Pakistan as visiting faculty member. His major areas of interest are Bayesian Statistics, mixture models, sample survey, descriptive statistics, probability distributions and income inequality. He is also serving many national and international journals as reviewer.

**Akhtar Nadeem** is an Associate Professor at the Department of Higher Education Archives and Libraries, Government of Khyber Pakhtunkhwa, Pakistan. He has specialized in the field of Bayesian Estimations, Categorical Data Analysis, Censored Data and Regression.

**Amin Muhammad** is working as Principal Scientist (Statistics). He has vast experience of data analysis in multidisciplinary research area in general and in food and agricultural research in particular. He has published more than 130 research papers in international/national journals and also presented his research work in many international/national conferences. He has completed his PhD in Probability and Mathematical Statistics and his thesis was Penalize quantile estimation and variable selecting in high dimensional data. He is a member of well-known statistical organizations.

**Amjad Ali** is an Associate Professor, Department of Statistics, Islamia College Peshawar, Pakistan. He has supervised many students and has published a number of research articles in national and international well reputed journals.

**Bhattacharjee Atanu** is working as a Lecturer in Medical Statistics at the University of Leicester, United Kingdom. He previously served as an Assistant Professor at the Section of Biostatistics, Centre for Cancer Epidemiology, Tata Memorial Centre, India, and the Malabar Cancer Centre, Kerala, India. He has completed his PhD at Gauhati University, Assam, on Bayesian Statistical Inference. He is an elected member of the International Biometric Society (Indian Region).

**Białek Jacek** has been an employee of Statistics Poland since 2018, where he is an expert in the Department of Trade and Services. Working in this position, he is involved in the analysis of scanner data and their application in the measurement of inflation. At the same time, Jacek Białek is an Associate Professor at the University of Lodz in Poland, where he works in the Department of Statistical Methods. Jacek Bialek's research interests revolve around the theory and practice of price indices. He is the author of more than 90 papers on price index theory, financial mathematics and open pension funds.

**Boumahdi Mounir** is a PhD student at the National School of Applied Sciences, faculty of Sciences Semlalia, University Cadi Ayyad, Complex Systems Modeling Laboratory. His main areas of interest include: statistical analysis of functional random variables, functional data analysis, big data, dernel estimator, surrogate response, surrogate data.

**Chugaievska Svitlana** is a Doctor of Economics at the Department of Economics and Innovation of the Jagiellonian University in Krakow, Poland and PhD in Economics, Associate Professor at the Department of Mathematical Analysis, Business Analysis and Statistics at Zhytomyr Ivan Franko State University, Ukraine. She is the Head of the Research Center for Socio-Economic Research at her Ukrainian University. He has 151 scientific and methodical publications, is a co-author of 5 textbooks/monographs on statistical modeling, economic analytics and macroeconomic analysis. Her area of scientific interest is sustainable development of economic entities of Ukraine in the context of European integration, transformation processes of social and demographic policy in Ukraine, economic challenges on the background of political stability/instability in the country: statistical assessment and analytics.

**Dehnel Grażyna** is an Assoc. Professor and Chair of the Department of Statistics, Poznań University of Economics and Business. Her main research domain is small area estimation, classification and data analysis methods, survey sampling, short-term and structural business statistics. She is also interested in outlier robust regression applied on business data and data integration. She is an author of numerous international and regional publications, including 2 books.

**Dey Rajashree** is a research scholar in the Mathematics and Computer Science division at the Institute of Advanced Study in Science and Technology (IASST), currently pursuing her PhD from the Academy of Scientific and Innovative Research (AcSIR), India. She holds a Post Graduate Diploma in Biostatistics from Tata Memorial Centre, India, and an MSc degree in Statistics from the Central University of Rajasthan, India. Her area of research interest is in biostatistics and epidemiology.

**Gershunskaya Julie** is a mathematical statistician with the Statistical Methods Staff of the Office of Employment and Unemployment Statistics at the U.S. Bureau of Labor Statistics. Her main areas of interest include statistical data integration, small area estimation, and treatment of influential observations, with application to the U.S. Current Employment Statistics Program.

**Grzenda Wioletta** is an Associate Professor at the Institute of Statistics and Demography at SGH Warsaw School of Economics. She has PhD in Mathematics. She has received habilitation in economics and finance from SGH Warsaw School of Economics for her works on Bayesian modelling of family and occupational careers. She has published papers on the applications of Bayesian and classical statistical methods in the analysis of unemployment and fertility and probability theory. She is an author and co-author of books on Bayesian statistics, advanced statistical methods, and programming in data analytics.

**Jimoh Toheeb A.** is a current doctoral student at the Department of Mathematics and Statistics, University of Limerick, Ireland through the Science Foundation Ireland Centre for Research Training in Foundations of Data Science. His current research interest encompasses statistics, machine learning, deep learning and natural language processing (NLP), specifically, text analytics, sentiment analysis and low-resource languages NLP. He is also an alumnus of the African Institute for Mathematical Science (AIMS), Rwanda.

**Kalton Graham** has wide-ranging interests in survey methodology, with a particular interest in survey sampling. He has been a research professor at the University of Maryland's Joint Program in Survey Methodology since the program was founded in 1993. In 2019 he retired from his position as a Senior Vice President at Westat, where he had worked for 27 years. His earlier positions were as a research scientist at the Survey Research Center and Professor of Biostatistics at the University of Michigan, the Leverhulme Professor of Social Statistics at the University of Southampton, and Reader in Social Statistics at the London School of Economics. He is a recipient of the Jerzy Neyman medal from the Polish Statistical Association, and he is a member of the Editorial Board for Statistics in Transition.

**Khan Sajjad Ahamad** is an Associate Professor, Department of Statistics, Islamia College Peshawar, Pakistan. He has supervised many students and has published number of research articles in national and international well reputed journals.

**Lahiri Partha** is Professor and Director of the Joint Program in Survey Methodology (JPSM) and Professor of Department of Mathematics at the University of Maryland College Park (UMD), and an Adjunct Research Professor of the Institute of Social Research, University of Michigan, Ann Arbor. Prior to joining UMD, Dr. Lahiri was the Milton Mohr Professor of Statistics at the University of Nebraska-Lincoln. His research interests include survey statistics, Bayesian statistics, data integration, and small-area estimation. He published over 80 papers in peer-reviewed journals, delivered 17 plenary/keynote speeches and over 80 invited talks in professional meetings worldwide. Over the years, Dr. Lahiri served on the editorial board of several international journals, including the Journal of the American Statistical Association and Survey Methodology. He served on several advisory committees, including the U.S. Census Advisory committee and U.S. National Academy panel, and served as consultant for international organizations such as the United Nations and the World Bank. Dr. Lahiri is a Fellow of the American Statistical Association and the Institute of Mathematical Statistics and an elected member of the International Statistical Institute. He received the 2021 SAE Award at the 63rd World Statistics Congress Satellite Meeting on Small Area Estimation in recognition of lifetime contributions to small area estimation research. More recently, Dr. Lahiri was awarded the Neyman Medal at a joint session of the 3rd Congress of Polish Statistics and 2022 International Association of Official Statistics (IAOS) held in Krakow, Poland, for outstanding contributions to the development of statistical sciences.

**Manzoor Sadaf** is an Assistant Professor, Department of Statistics, Islamia College Peshawar, Pakistan. He has supervised many students and has published a number of research articles in national and international well reputed journals.

**Melesse Sileshi Fanta** is an associate Professor in School of Mathematics, Statistics and Computer Science at the University of KwaZulu Natal (UKZN) South Africa. He holds three MSC degrees in different areas of Statistics and a PhD in Statistics from UKZN. His publications include issues related to are Linear Mixed models, Nonlinear Mixed Models, Longitudinal Data Analysis, Survival Analysis, Spatial modelling and Path modelling. He has published more than 45 research papers in international/national journals.

**Münnich Ralf** is Full Professor and Chairholder of the Economic and Statistics Department or Trier University. His main research areas include survey statistics, variance estimation, and statistical microsimulation methods. Ralf Münnich was responsible for several EU projects as well as for the German Census 2011 and 2022

sampling and estimation projects. Currently, he is speaker of the research unit FOR 2559 on microsimulation funded by the German Research Foundation. Since 2020, he has been chairman of the German Statistical Society.

**Ndlovu Bonginkosi Duncan** is a lecturer at the Durban University of Technology. His main area of interest is survival analysis with a particular focus on discrete time competing risks.

**Olalude Gbenga A.** is a Principal Lecturer at the Department of Statistics, School of Applied Sciences, Federal Polytechnic Ede, Osun State, Nigeria. His main areas of interest include: time series, Bayesian inference and mathematical statistics. He has published more than 50 research papers in international/national journals and conferences. He is a member of Nigerian Statistical Association, American Statistical Association, and Professional Statisticians Society of Nigeria.

**Olayinka Hammed A.** is a doctoral student and Teaching Assistant at the Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, Massachusetts, USA. His areas of research interest include: high frequency financial time series, time series econometrics, biostatistics, spatial statistics, high dimensional statistical inference, and Bayesian methodologies for risk analysis. He has 7 publications/conference papers to his credit. Currently, he is a student member of Mathematical Association of America and American Mathematical Society.

**Öztaş Ayhan H.** is Professor Emeritus in the Department of Statistics at Middle East Technical University, Ankara, Turkey. His research interests are survey sampling techniques, survey methodology research, and web survey methodology. During the 1990-93 period, he served as the Presidential Adviser and Director of the Technical Services Department of the State Institute of Statistics Turkey. He has been an elected extraordinary member of the International Statistical Institute since 1995. He has been the Country Representative for Turkey of the International Association of Survey Statisticians since 1989. Also, Vice President of Turkish Statistical Association (2007–2009). Professor Ayhan has published 17 books/monographs and more than 100 research papers in international/national journals and conferences. Professor Ayhan is an active member of the ISI, IASS, IAOS, TSA, and TPA.

**Pfeffermann Danny** is Professor Emeritus at the Hebrew University, Israel, and Professor at Southampton University, UK. He served for 9 years as the National Statistician and Director General of the CBS of Israel. His main research areas are analytic inference from complex sample surveys, small area estimation, seasonal adjustment and trend estimation, non-ignorable nonresponse, mode effects and proxy surveys. Professor Pfeffermann published more than 80 articles in refereed journals and co-edited the two-volume handbook “Sample Surveys”, published by North-Holland. He served as Ass. Editor of several journals. Professor Pfeffermann was President of the

Israel Statistical Society and of the International Association of Survey Statisticians (IASS). He is a Fellow of the American Statistical Association and the Institute of Mathematical Statistics, and an elected member of the International Statistical Institute.

**Rachdi Mustapha** is currently working as a Full Professor at the Department of Mathematics and Computer Sciences with Applications to Social Sciences, Alps Grenoble University (France). He obtained his Ph.D. in Statistics from Rouen University (France) and his Habilitation to Direct Researches (HDR) from Alps Grenoble University (France). His Areas of Research Interests are functional and Complex Data Analyses, Applied Statistics, and Statistical Deep Learning. Published more than 120 articles in refereed journals.

**Sharma Anand** is an Associate Professor of Economics at Jindal Global Business School, O.P. Jindal Global University, Sonapat, India. He is a Fellow of the Indian Institute of Management, Ahmedabad. Previously, Dr. Sharma was associated with IIM Rohtak and IIM Sirmaur. He has also worked as an economic expert in the Antitrust Division of the Competition Commission of India. His research interests include development economics, health economics, and institutional economics. He has published research papers in leading national and international journals.

**Sharma Vipin** is an Associate Professor at the University School of Business-APEX-MBA, Faculty of Economics, Chandigarh University, Punjab, India. He also serves as the department coordinator of AMBA, NBA, NAAC, IQAC and Research. His main areas of interest include intra-regional trade, econometrics, international trade, financial inclusion, and economics of education. Dr. Vipin Sharma has published more than 17 research papers in international/national journals and conferences and 5 five book chapters. He has delivered 8 lectures as a resource person/keynote speaker in FDPs, workshops, debate competitions, and conferences held at the national and international levels.

**Sharma Vishwantra** is a Senior Consultant at the Directorate of Census Operations, Jammu, J&K, India. She received the PhD degree in Statistics from University of Jammu, J&K, India. Her main areas of interest include: Sample survey, Estimation Theory and Statistical Inference. She is also a reviewer for various esteemed journals in the area of mathematics and statistics.

**Targonskii Andrey** is a PhD in Mathematics, Associate Professor of the Department of Mathematical Analysis, Business Analysis and Statistics of Zhytomyr Ivan Franko State University, Ukraine. The sphere of his scientific interests is applied mathematics, stochastic and complex analysis, geometric theory of a complex variable functions. He has more than 50 publications in international journals and speeches at scientific conferences. He is the author of 3 educational manuals on mathematical analysis.



**Tiwari Kuldeep Kumar** is an Assistant Professor at the Department of Mathematics, University Institute of Sciences, Chandigarh University, Mohali, India. He received his PhD from Shri Mata Vaishno Devi University, Jammu, India. His main areas of interest include: sampling survey, estimation theory and information theory. He is also a reviewer for various esteemed journals in the area of mathematics and statistics.

**Tokas Shekhar** is an Assistant Professor at the School of Global Affairs, Dr. B. R. Ambedkar University, Delhi, India. He has a PhD in Economics from Jawaharlal Nehru University and completed his Masters and Bachelors in Economics from Jamia Millia Islamia and Hansraj College, University of Delhi, respectively. He has held lectureships for more than six years at the University of Delhi and Delhi Technological University. His research and publications have primarily focused on the international migration of students and knowledge workers, human capital investment, internationalisation of higher education, and development economics. Dr. Tokas is also the recipient of the 'Major Research Grant' awarded by the Indian Council for Social Science Research (ICSSR) and O.P. Jindal Global University (2023–2025), and ICSSR Doctoral Fellowship (2014–2015).

**Vogt Martin** is a Full Professor for Business Intelligence, in particular Advanced Analytics at the Business School at Trier University of Applied Sciences. Prior to that, Prof. Vogt accumulated extensive experience in the financial industry, holding various management positions. His main areas of interest include natural language processing, data analysis and risk management. His research has received several awards, including the Gerhard Fürst Prize, awarded by the German Federal Statistical Office.

**Wesołowski Jacek** is a Professor of mathematics. His research interests include wide range of subjects within probability, stochastic processes, mathematical statistics and survey methodology. He has published more than 170 research papers, mostly in prestigious probabilistic journals including such as: *Annals of Probability*, *Probability Theory and Related Fields*, *Annals of Statistics*, *Transactions of the American Mathematical Society*, *Journal of Functional Analysis*, *Survey Methodology*. He is also an author (jointly with G. Rempala) of a Springer monograph on random matrices. Professor Wesołowski is professionally affiliated with Warsaw University of Technology and with Statistics Poland.

**Yaya OlaOluwa S.** (PhD) is a senior lecturer at the University of Ibadan, Nigeria and University Professor & Doctoral advisor at the Global Humanistic University (GHU), Curacao. He has interest in time series econometrics, economic modelling, and data science. He holds various international and local research fellowships such as Research Fellowships from the Navara Centre for International Development, University of Navara, Spain; University of Ho Chi Minh City, Vietnam; ILMA University, Karachi, Pakistan; KMU Akademie, Austria; Humanistic University, Curacao, among others.

He is a two-time winner of Prof Adenike Osofisan Distinguished Science Faculty Scholar award. He is the Deputy Director & Research Fellow Centre for Econometrics and Applied Research (CEAR), Ibadan, Nigeria. He has over 100 publications and two books to his credit, with majority listed in ISI/Scopus database.

**Zewotir Temesgen** is a Professor of Applied Statistics and Data Science at the University of KwaZulu Natal, South Africa. He is a co-director of Data Science Unit for Business and Industry. He has a proven strong track record in research and supervision of postgraduate studies. He has published more than 150 research articles

# GUIDELINES FOR AUTHORS

We will consider only original work for publication in the Journal, i.e. a submitted paper must not have been published before or be under consideration for publication elsewhere. Authors should consistently follow all specifications below when preparing their manuscripts.

## Manuscript preparation and formatting

The Authors are asked to use *A Simple Manuscript Template (Word or LaTeX) for the Statistics in Transition Journal* (published on our web page: <https://sit.stat.gov.pl/ForAuthors>).

- **Title and Author(s).** The title should appear at the beginning of the paper, followed by each author's name, institutional affiliation and email address. Centre the title in **BOLD CAPITALS**. Centre the author(s)'s name(s). The authors' affiliation(s) and email address(es) should be given in a footnote.
- **Abstract.** After the authors' details, leave a blank line and centre the word **Abstract** (in bold), leave a blank line and include an abstract (i.e. a summary of the paper) of no more than 1,600 characters (including spaces). It is advisable to make the abstract informative, accurate, non-evaluative, and coherent, as most researchers read the abstract either in their search for the main result or as a basis for deciding whether or not to read the paper itself. The abstract should be self-contained, i.e. bibliographic citations and mathematical expressions should be avoided.
- **Key words.** After the abstract, Key words (in bold) should be followed by three to four key words or brief phrases, preferably other than used in the title of the paper.
- **Sectioning.** The paper should be divided into sections, and into subsections and smaller divisions as needed. Section titles should be in bold and left-justified, and numbered with 1., 2., 3., etc.
- **Figures and tables.** In general, use only tables or figures (charts, graphs) that are essential. Tables and figures should be included within the body of the paper, not at the end. Among other things, this style dictates that the title for a table is placed above the table, while the title for a figure is placed below the graph or chart. If you do use tables, charts or graphs, choose a format that is economical in space. If needed, modify charts and graphs so that they use colours and patterns that are contrasting or distinct enough to be discernible in shades of grey when printed without colour.
- **References.** Each listed reference item should be cited in the text, and each text citation should be listed in the References. Referencing should be formatted after the Harvard Chicago System – see <http://www.libweb.anglia.ac.uk/referencing/harvard.htm>. When creating the list of bibliographic items, list all items in alphabetical order. References in the text should be cited with authors' name and the year of publication. If part of a reference is cited, indicate this after the reference, e.g. (Novak, 2003, p.125).